An Intelligent Banking Decision Making Model to Optimize Banks Creditworthiness

Sherif ELsaied Elsaied Gad[1], Nashaat ELkhameesy[2], Nevine Makram Labib[3]

Ph.D. Candidate Department of Computer and Information Systems Sadat Academy for Management Sciences, Cairo, Egypt.

Email: doctorsherif2021@gmail.com

Professor of Computer and Information Systems, Sadat Academy for Management Sciences, Cairo, Egypt.

Email: wessasalsol@gmail.com

Professor of Computer and Information Systems, Sadat Academy for Management Sciences, Cairo, Egypt.

Email: nevmakram@gmail.com

Abstract

Effective banking decision-making is crucial for financial institutions to optimize customer engagement and risk management. Data mining techniques, particularly Support Vector Machines (SVM), offer powerful predictive capabilities for enhancing decision-making processes. This study applies SVM to a benchmark dataset of banking institutions› direct marketing campaigns to predict customer acceptance of term deposit offers. The methodology involves data preprocessing, feature analysis, and model optimization to handle class imbalances and improve classification accuracy. The dataset was initially split into training and testing. However, due to the imbalance in the targeted data set, we adopted the balanced class weight approach and 10-fold cross-validation to solve the imbalanced class problem. Experimental results demonstrate that SVM effectively classifies potential customers, improving the accuracy of marketing decisions. Moreover, the results justified the potential of SVM in enhancing banking decision-making by optimizing telemarketing strategies and increasing customer conversion rates. The proposed model outperformed the counterpart approaches in terms of performance measures and achieved perfect classification.

Keywords: Machine Learning; Bank Deposit Prediction; Support Vector Machines (SVM); Bank Telemarketing; Data Mining

1. Introduction

Banking institutions play a crucial role in financial stability and economic growth by making strategic decisions on credit approvals, risk management, and customer engagement. Effective decision-making is essential for banks to enhance profitability and reduce financial risks. With the increasing volume of economic data, traditional decision-making approaches often struggle to extract meaningful insights [1].

Customers deposit money in banks to earn annual interest, while banks use these deposits to provide loans to diverse clients. Banks› profit comes from the

difference between the interest rates on deposits and loans. Banks play a crucial role in economic development by offering deposit services, as these funds support business loans and investments that drive national growth. Consequently, bank deposits contribute significantly to the financial infrastructure supporting economic progress. Additionally, customer databases contain key attributes that influence a depositor's potential. However, manually analyzing this data to identify reliable depositors is challenging. The process can be streamlined by recognizing the most relevant input characteristics [2].

Telemarketing, a direct marketing strategy, involves sales representatives contacting potential customers via phone to promote products or services [3]. In the banking sector, bank telemarketing is widely used to promote financial products such as term deposits, credit cards, and loan offers. Banks rely on these campaigns to attract new clients, cross-sell services, and retain existing customers in a highly competitive market [4]. Many financial institutions have adopted telemarketing to attract new customers, enhance existing services, and meet client needs. However, direct marketing efforts are unsuccessful, as consumers often prefer well-established financial institutions [5]. The advancement of telemarketing through digital and mobile technology has simplified the process of generating reports and analyzing marketing campaign outcomes.

Therefore, advanced data mining techniques have become essential for improving banking decision-making by identifying patterns, predicting customer behavior, and optimizing marketing strategies [6].

Machine learning has become an essential component of artificial intelligence (AI), enabling systems to perform complex tasks such as forecasting, planning, analysis, recognition, and robot control [7]. It focuses on developing algorithms that can predict data patterns and improve performance over time by adapting to previous inputs. By leveraging customized tuning settings, machine learning enhances the efficiency of systems, mimicking human cognitive processes to handle multi-level data representations while addressing the selectivity-invariance challenge [8].

In the financial sector, particularly in banking, machine learning is crucial in optimizing customer interactions and decision-making [9]. Telemarketing remains a primary method for banks to reach potential clients, and machine learning can automate feature validation processes to determine customer availability for direct marketing campaigns. However, conventional approaches assign varying weights to different factors, whereas real-world banking decisions may often be influenced by a single dominant factor, which these models may not fully capture.

Data mining, a subset of artificial intelligence, enables banks to analyze vast amounts of structured and unstructured data to support decision-making processes. Techniques such as classification, clustering, and predictive modeling help in areas like credit scoring, fraud detection, and customer segmentation [10]. Support Vector Machines (SVM) have gained prominence among these techniques due to their ability to handle high-

dimensional data and achieve robust classification performance [11].

Several data mining techniques, such as the One-R Algorithm, Naïve Bayes (NB) classifier, classification models, and association rule mining, have been utilized in banking direct marketing to enhance service categorization [12]. After performing exploratory data analysis to identify relationships between variables and outcomes, data mining techniques categorize bank customers based on their likelihood of subscribing to term deposits. The primary objective of classification models is to predict whether a customer will accept a term deposit offer. This involves developing a predictive model based on training data, where class labels are known, to classify new data points with unknown labels accurately.

By integrating machine learning with banking decision-making, financial institutions can improve customer targeting, enhance marketing efficiency, and optimize overall business strategies. This study explores the application of machine learning techniques, particularly Support Vector Machines (SVM), in banking decision-making to enhance predictive accuracy in direct marketing campaigns. This study applies Machine Learning (ML) to a benchmark dataset from financial institutions› direct marketing campaigns to develop an accurate classifier for predicting customer acceptance of long-term deposit offers. The objective is to develop an accurate predictive model that classifies potential customers for term deposit acceptance, ultimately enhancing targeted marketing efforts and customer acquisition strategies. The research highlights the significance of machine learning in banking and provides insights into improving financial decision-making through advanced data analytics. This is particularly important in the context of bank telemarketing, where the ability to accurately identify and engage likely responders can substantially increase the efficiency and return on marketing investments. The findings highlight the significant impact of machine learning techniques on telemarketing campaign outcomes. The research follows two key phases: data preparation and model evaluation. In the data preparation phase, preprocessing involves cleaning the data by removing duplicates and handling missing values. Data visualization is then performed before encoding categorical variables using label and one-hot encoding techniques.

The dataset was initially split into training and testing sets. However, due to class imbalance, adjustments were made during training using a balanced class weight strategy and 10-fold cross-validation to improve model performance. The SVM algorithm was employed for training and testing, resulting in a highly accurate classifier. The proposed model achieved perfect classification, outperforming all existing state-of-the-art model.

The remainder of this paper is structured as follows: Section 2 presents related studies that have utilized the same dataset. Section 3 details the proposed model, including dataset preprocessing, analysis, visualization, encoding techniques, machine learning-based prediction, and the results achieved. Finally, Section 4 provides the conclusion of the study.

2. Related Work

Several studies have explored the application of various ML techniques in predicting the effectiveness of bank telemarketing for long-term deposit sales. Recently, applying ML in banking decision-making has gained a significant attention as it enables financial institutions to enhance their customer engagement and decision-making processes. Many studies have explored machine learning techniques to improve marketing strategies, risk management, and customer segmentation in the banking industry. Below, is a brief review on some recent research literatures that have targeted the bank marketing dataset.

One prominent Area of research is using machine learning for direct marketing, specifically telemarketing campaigns. Several studies have focused on predicting customer behavior and improving campaign effectiveness. For example, in their research, Lacerda et al. [13] applied ML-models, including Decision Trees and Random Forests, to the bank marketing dataset to predict customer acceptance of term deposits. Their results demonstrated that ML- techniques significantly enhance the targeting of potential clients, leading to better marketing outcomes. Similarly, Moreira et al. [14] employed Support Vector Machines (SVM) and neural networks to classify customers based on their likelihood of responding to marketing offers. The study showed that machine learning algorithms outperformed traditional statistical methods regarding predictive accuracy and campaign efficiency.

In addition to marketing, machine learning has been applied to other banking decision-making areas, such as credit scoring and fraud detection. Peter et al. [15] implemented ensemble learning approaches–namely bagging, boosting, and stacking–to improve the precision and robustness of predicting customer subscription behavior in bank telemarketing campaigns. Acknowledging the difficulties associated with class imbalance and the intricacies of customer behavior, they utilized ensemble methods to construct a more resilient prediction model. Their results indicate that stacking outperforms other models, achieving an accuracy of 91.88% and an Area Under the Curve (AUC) score of 0.9491. Furthermore, the feature importance analysis highlighted that variables such as call duration, macroeconomic indicators like the Euribor rate, and customer age significantly influence subscription outcomes. These insights emphasize the potential of leveraging customer interaction data and economic conditions to enhance the effectiveness of telemarketing strategies in the banking sector.

Aibden et al. [16] introduced a hybrid methodology that integrates SVM with the Synthetic Minority Oversampling Technique (SMOTE) and hyperparameter optimization using GridSearchCV to improve predictive performance. The preprocessing pipeline involved categorical variable encoding, feature scaling, and addressing class imbalance through SMOTE-based resampling. The resulting optimized SVM

model achieved an accuracy of 91% and an AUC of 0.96, demonstrating its effectiveness in predicting customer responses to term deposit offers.

Tanvir et al. [17] investigated the performance of logit and probit regression models for predicting customer subscriptions to term deposits, utilizing a direct marketing dataset from a Portuguese bank. The dataset comprises various demographic, economic, and behavioral attributes that influence the likelihood of subscription. To address class imbalance and enhance model reliability, the target variable was balanced prior to training. Bayesian methods and Leave-One-Out Cross-Validation (LOO-CV) were employed to assess the predictive capabilities of both models. The findings revealed that the logit model outperformed the probit model in effectively managing the classification task. Sarlak et al. [18], machine learning models such as Random Forest and Gradient Boosting were used to predict customer churn, providing insights into customer retention strategies and decision-making.

The classifier developed by Safarkhani and Moro [19], accurately predicted which customers would accept a bank›s long-term deposit offer. Using the bank marketing dataset, they focused on resampling to address class imbalance, feature selection to simplify data processing, and dimension reduction to eliminate inefficiencies in the modeling process. Their approach resulted in improved classification algorithm performance. The experimental results showed that the J48 decision tree achieved a prediction accuracy of 94.39%, with a sensitivity of 0.975 and specificity of 0.709.

Borugadda et al. [20] explored the effectiveness of telemarketing techniques in promoting long-term bank deposits among potential clients. The study applied several machine learning methods, including Random Forest (RF), SVM, Gaussian Naive Bayes (GNB), Decision Tree (DT), and Logistic Regression, to analyze the demand for long-term bank deposits (LR) using the bank marketing dataset. The results revealed that the LR model achieved an accuracy of 92.48%. Additionally, the findings provide valuable insights for banks in making telemarketing policy decisions, helping them assess deposit offerings› success to existing and prospective clients.

Phung and Khuat [21] demonstrated how machine-learning approaches can influence the outcomes of telemarketing campaigns. The study consisted of two main phases: data preparation and model evaluation. In the first phase, data cleaning involved removing duplicate records, addressing missing values, visualizing the dataset to assess balance, and applying a response coding method with Laplace smoothing to encode categorical data. Additionally, including a «duration» variable impacted the results significantly. In the second phase, various machine learning models were evaluated to select the best classifier, including K-Nearest Neighbors (KNN), Logistic Regression (LR), Linear SVM, and Extreme Gradient Boosting (XGBoost). Due to data imbalance, the AUC score

was used to assess model performance. The results showed that KNN was the most accurate, achieving a 91.07% accuracy and a 93% AUC. The study›s findings suggest that using KNN with a k-value greater than five offers valuable insights from a commercial perspective.

Zaki et al. [22] aims to predict customer subscriptions to bank term deposits. The evaluation considered several classifiers–Stochastic Gradient Descent, k-nearest neighbors, and Random Forest. Of these, the Random Forest model delivered the strongest results, achieving 87.5% accuracy, a negative predictive value of 92.9972%, and a positive predictive value of 87.8307%. Abd et al. [23] employed several machine learning algorithms–DT, KNN, CatBoost, and RF–to classify customer data with the objective of enhancing the prediction of customer behavior and facilitating more efficient business engagement. The system was evaluated using the same dataset. Among the models, Random Forest achieved promising accuracy at 0.97, followed by Decision Tree with 0.95, and KNN with 0.91. Notably, the CatBoost algorithm, with an execution time of 15.04 seconds, yielded the best overall performance, attaining the highest F1 score of 0.91 and accuracy of 0.98.

Overall, the growing body of research demonstrates the effectiveness of machine learning in various facets of banking decision-making, from enhancing marketing campaigns to improving risk management and customer segmentation. Applying machine learning techniques, particularly classifiers like SVM, Random Forest, and Decision Trees, has significantly improved predictive accuracy, making them valuable tools for modern banking practices.

Among these, SVM is especially suitable for this study due to their strong performance in handling high-dimensional feature spaces, their robustness to overfitting in cases with fewer samples relative to features, and their ability to construct non-linear decision boundaries using kernel functions. These characteristics make SVM highly effective for binary classification problems such as predicting customer acceptance in telemarketing campaigns.

3. The Proposed Model

This section outlines the proposed model, detailing the dataset and the flow work stages of the adopted methodology which include: preprocessing, analysis, and visualization tasks. It also explains the encoding process, the application of machine learning for prediction, and the results achieved using the proposed approach.

3.1. Dataset

The dataset used in this study is sourced from Kaggle [24] and consists of 45,211 records, each containing 17 attributes relevant to Portugal›s bank marketing campaigns. These attributes represent key factors influencing campaign success. The target variable indicates whether a customer subscribes to a term deposit, making this a binary classification problem. Table 1 details the dataset's attributes, including their types and roles.

Table 1: Dataset Characteristics

Attribute	Type	Description	Values Present
Age	Numeric	Customer's Age	18 - 95
Job	Categorical	Customer's Profession	‹management,› ‹technician,› ‹entrepreneur›, ‹blue›, ‹unknown›, ‹retired›, ‹admin,› ‹services,› ‹self›, ‹unemployed,› ‹housemaid›, ‹student›
Marital	Categorical	Marital Status of Customer	‹married,› ‹single›, ‹divorced›
education	Categorical	Education Qualification	‹tertiary,› ‹secondary,› ‹unknown›, ‹primary›
Default	Categorical	Whether the customer has credit in default	‹no›, ‹yes›
Balance	Numeric	Balance present in the account	-8019 -102127
Housing	Categorical	Whether the customer has a housing loan or not	'yes', 'no'
Loan	Categorical	Whether the customer acquires a housing loan or not	'no', 'yes'
Contact	Categorical	contact communication type	'unknown,' 'cellular,' 'telephone'
Attribute	Type	Description	Values Present
Day	Numeric	last contact day	1- 31
Month	Categorical	last contact month of the year	‹may›, ‹jun›, ‹jul›, ‹aug›, ‹oct›, ‹nov›, ‹dec›, ‹jan›, ‹feb›, ‹mar›, ‹apr›, ‹sep›
Duration	Numeric	last contact duration, in seconds	0- 4918
Campaign	Numeric	number of contacts performed during this campaign and for this client	1- 63
Pdays	Numeric	number of days that passed by after the client was last contacted from a previous campaign	1 - 871
Previous	Numeric	number of contacts performed before this campaign and for this client	0- 275
Poutcome	Categorical	the outcome of the last marketing campaign	'unknown,' 'failure,' 'other', 'success'
Y	Numerical	Whether the client subscribed to a term deposit or not	0,1

3.2. Preprocessing

We examined the dataset for missing or inconsistent values across all features. Specifically, a small number of entries labeled as «unknown» were identified in the employment, education, and interaction attributes. These entries were treated as missing values and were handled using mode imputation [25], replacing each «unknown» with the most frequent category in its respective feature to preserve categorical integrity. To ensure value consistency, we performed range checks [26] on numerical attributes and category validation on nominal variables by comparing observed categories to the predefined schema (e.g., valid education levels, employment statuses).

To mitigate multicollinearity and reduce the risk of overfitting, we computed the Pearson Correlation Coefficient [27] between all pairs of independent variables. Correlation values were interpreted as follows: 0 to 0.3 indicates a weak relationship, 0.3 to 0.7 a moderate one, and 0.7 to 1 a strong correlation. As shown in Figure 1, the correlation matrix confirms the absence of strong linear relationships among the features, suggesting that the dataset does not contain redundant predictors.

	age	balance	day	duration	campaign	pdays	previous
age	1.000000	0.097783	-0.009120	-0.004648	0.004760	-0.023758	0.001288
balance	0.097783	1.000000	0.004503	0.021560	-0.014578	0.003435	0.016674
day	-0.009120	0.004503	1.000000	-0.030206	0.162490	-0.093044	-0.051710
duration	-0.004648	0.021560	-0.030206	1.000000	-0.084570	-0.001565	0.001203
campaign	0.004760	-0.014578	0.162490	-0.084570	1.000000	-0.088628	-0.032855
pdays	-0.023758	0.003435	-0.093044	-0.001565	-0.088628	1.000000	0.454820
previous	0.001288	0.016674	-0.051710	0.001203	-0.032855	0.454820	1.000000

Figure 1: Correlation Matrix

3.3. Data Exploration and Visualization

The dataset features can be categorized as either numerical or categorical. Numerical attributes include «age,» «balance,» «duration,» «campaign,» «pdays,» and «previous,» while categorical

features comprise «job,» «marital,» «education,» «default,» «housing,» «loan,» «poutcome,» and «y.» For categorical variables, bivariate analysis and visualizations are conducted to examine their relationship with the target variable «y.» Bivariate analysis is instrumental in helping contact centers identify specific consumer groups to target.

Figure 2 illustrates the correlation between the «job» feature and «y,» showing that banks approached more individuals with professional backgrounds, and highly educated individuals were more likely to subscribe to term deposits. Figure 3 depicts the relationship between «marital» status and «y,» revealing that married and single individuals were targeted more frequently than divorced ones, with married individuals having a higher subscription rate. Figure 4 highlights the connection between «education» and «y,» indicating that customers with higher education levels were more inclined to opt for term deposits, demonstrating a proportional relationship.

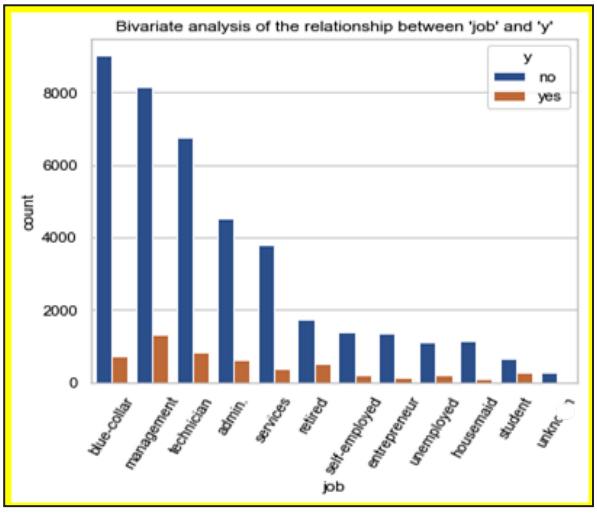

Figure 2: Relationship between job and y.

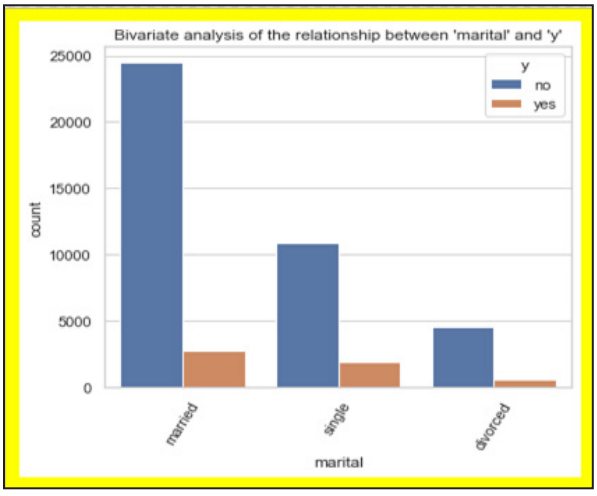

Figure 3: Relationship between marital and y.

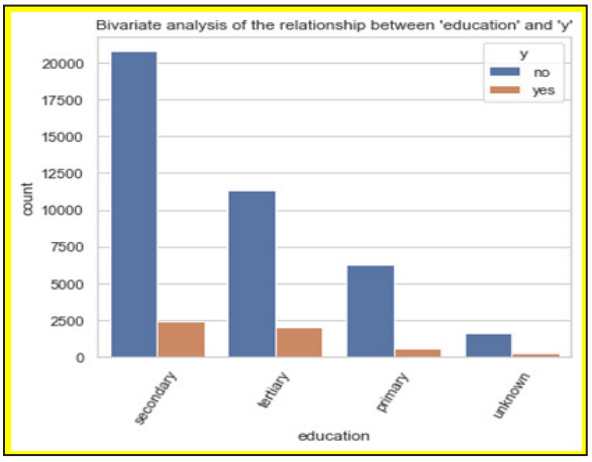

Figure 4: Relationship between education and y.

Additionally, Figure 5 shows the impact of the «default» feature on «y,» where non-defaulters had a higher subscription rate. It was observed that individuals with existing credit were less likely to subscribe to new bank offers, which aligns with expectations.

Figure 6 illustrates the relationship between the «housing» feature and «y,» showing that while most individuals have a mortgage, those without a mortgage were more likely to subscribe to a term deposit. This indicates an inverse relationship.

Figure 7 depicts the connection between the «loan»

feature and «y.» Like housing loans, individuals without personal loans showed a higher inclination toward subscribing to a term deposit. In contrast, only a small percentage of personal loan holders opted for a subscription, indicating a direct proportional relationship.

Figure 8 demonstrates the impact of the «contact» feature on «y.» The graph reveals that customers contacted via cellular phones had a higher enrollment rate for term deposits. Any «unknown» values in this category will be handled using an imputation approach.

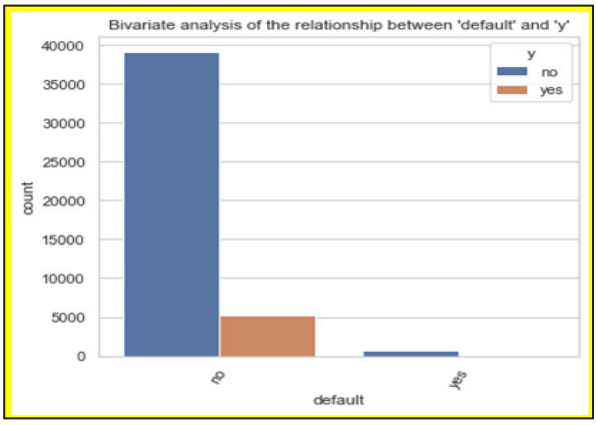

Figure 5: Relationship between default and y.

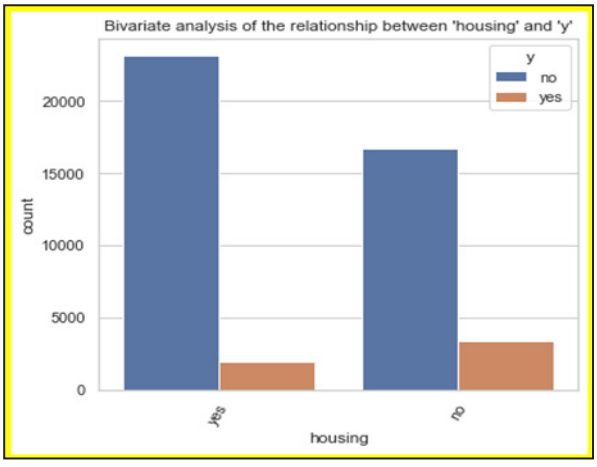

Figure 6: Relationship between housing and y

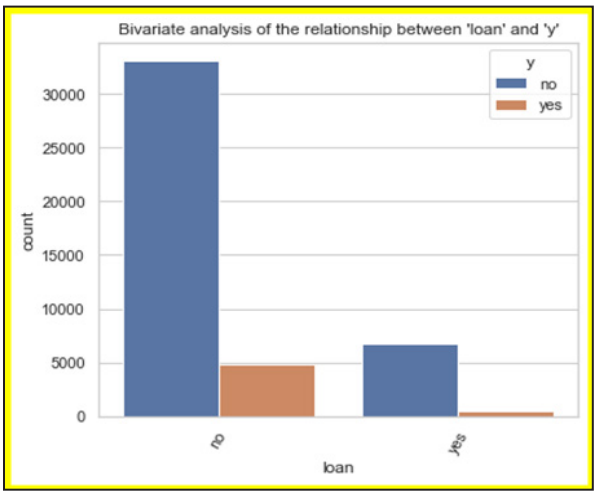

Figure 7: Relationship between loan feature and y

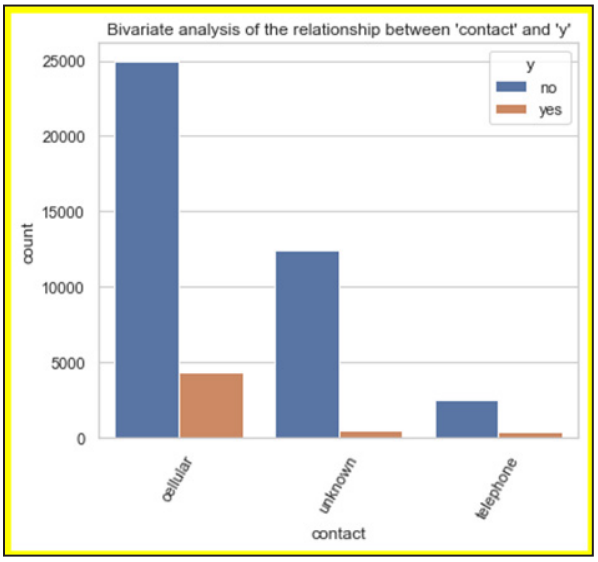

Figure 8: Relationship between contact and y.

Figure 9 explores the relationship between the «month» feature and «y,» showing a slight increase in subscriptions during May compared to previous months. Except for December and January, the subscription rate remains relatively stable despite fluctuations in the number of contacts. The decline in subscriptions during these months may be attributed to vacation periods. As the «month» feature does not significantly impact the outcome, it

will be excluded from further analysis. To rigorously evaluate its predictive contribution, we examined the feature importance scores derived from the SVM model and observed that month ranked among the least informative features. Additionally, removing month did not negatively affect model performance in terms of accuracy, precision, or recall. Therefore, this feature was excluded to reduce model complexity without compromising predictive power. Figure 10 examines the «poutcome» feature's relationship with «y,» aligning with the success of previous marketing campaigns. The «unknown» category, representing 78.7% of contacts, indicates that most individuals were unaware of prior campaigns. While the «success» category has a lower proportion, it remains valuable in understanding the campaign›s effectiveness.

Figure 11 presents the "univariate analysis" of the target variable «y.» Most individuals contacted declined the bank›s offer, with only 13.2% agreeing to subscribe. This substantial imbalance highlights a class distribution issue, which needs to be addressed in the analysis.

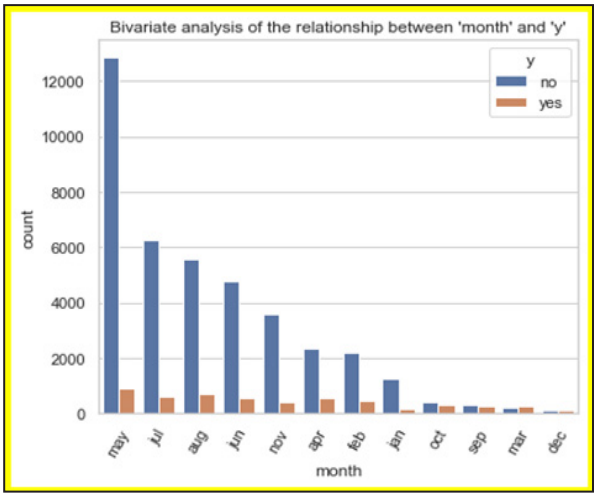

Figure 9: Relationship between month and y

Box plots and histograms [28] have been utilized to analyze numerical features, along with each attribute›s median, mean, and mode [29]. Figure 12 illustrates the analysis of the «age» feature, using box plots to depict its relationship with the classified target variable. The histogram presents a left-skewed normal distribution, forming a bell shape. The population›s age distribution ranges between 20 and 60 years, while the interquartile range in the box plot lies between 30 and 50 years. This trend likely reflects increased financial stability and productivity during this period. A similar pattern emerges when examining the «job» feature.

Figure 13 analyzes the «balance» feature, showing a median of 0, indicated by the green line. This suggests that most contacted individuals had an annual balance close to $0.

Figure 14 analyzes the «day» feature, where the histogram demonstrates a symmetrical pattern with a peak on the 20th day. However, this feature does not significantly influence the results, as individuals may subscribe on any given day. Therefore, it will be excluded from further analysis.

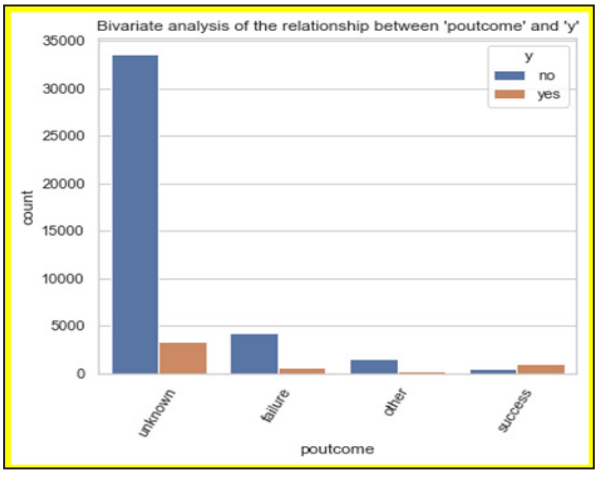

Figure 10: Relationship between poutcome and y

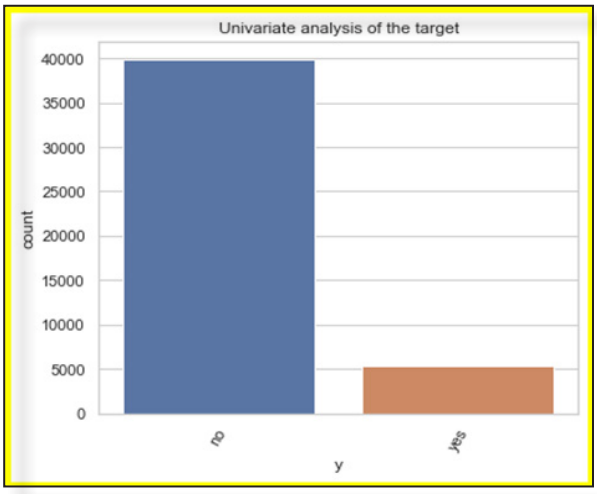

Figure 11: Univariate analysis for target y

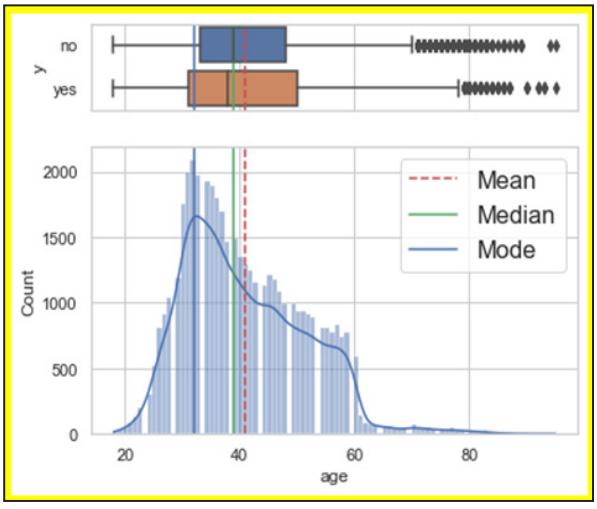

Figure 12: Age feature analysis

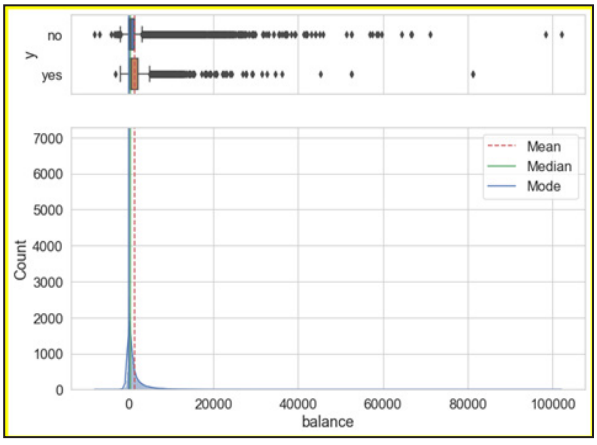

Figure 13: Balance feature Analysis.

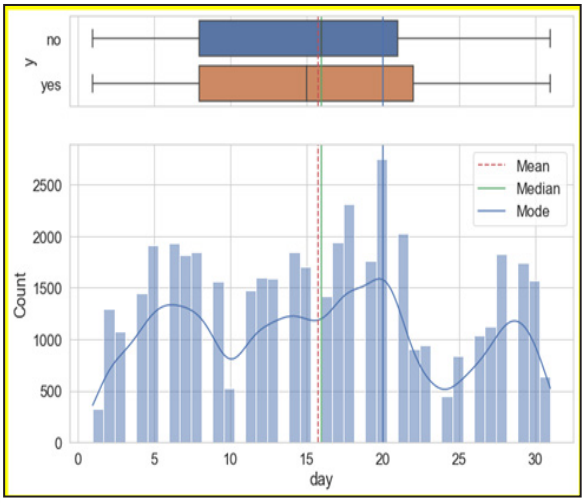

Figure 14: Day feature analysis.

Figure 15 examines the «duration» feature, significantly impacting the target variable «y.» The analysis indicates that most consumers reject the offer within the first two minutes of contact. However, the likelihood of acceptance increases for calls lasting between 2 to 12 minutes. Very few individuals take an extended time to decide on an offer.

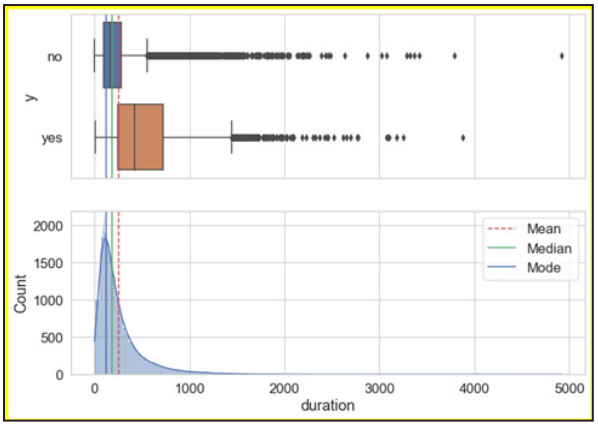

Figure 15: Duration feature analysis

Figure 16 presents an analysis of the «campaign» feature, representing the total number of contacts made during the campaign. The plots indicate that the distribution between «yes» and «no»

responses is relatively balanced. Additionally, more frequently contacted individuals were more likely to subscribe to a term deposit. However, excessive contact attempts appear inefficient and do not significantly improve subscription rates.

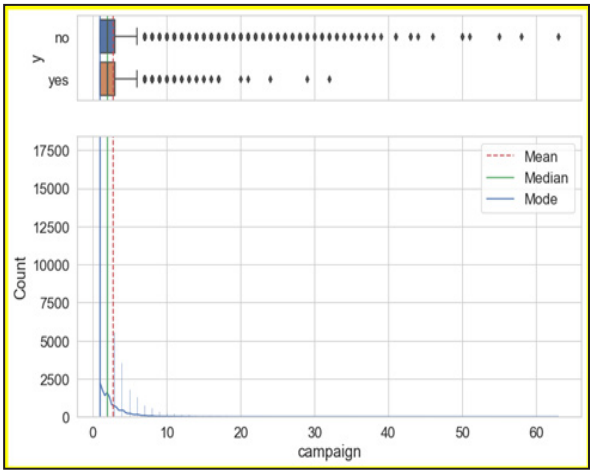

Figure 16: Campaign feature analysis.

Figure 17 examines the «pdays» feature, which records the number of days since a customer was last contacted in a previous campaign (where -1 indicates no prior contact). The box plot reveals that most individuals had never been approached before, with a median value of -1. Since this feature does not contribute meaningful insights to the prediction, it will be excluded from the analysis. This sparsity and lack of variance were reflected in the feature importance analysis, which ranked pdays as non-influential. Further, model evaluation with and without pdays showed no statistically significant difference in classification performance, justifying its removal.

Figure 18 analyzes the «previous» feature, which indicates the total number of contacts made with a customer before the campaign. The data shows that 36,954 individuals in this campaign were contacted for the first time. The absence of a box plot suggests no meaningful distribution or correlation with the target variable. Consequently, this feature will also be removed from the dataset. The SVM model assigned it minimal importance, and its exclusion did not degrade model metrics. Moreover, the lack of distributional variation and correlation with the target variable undermines its utility for classification. Therefore, to prevent noise and overfitting, the previous feature was also omitted from the final dataset.

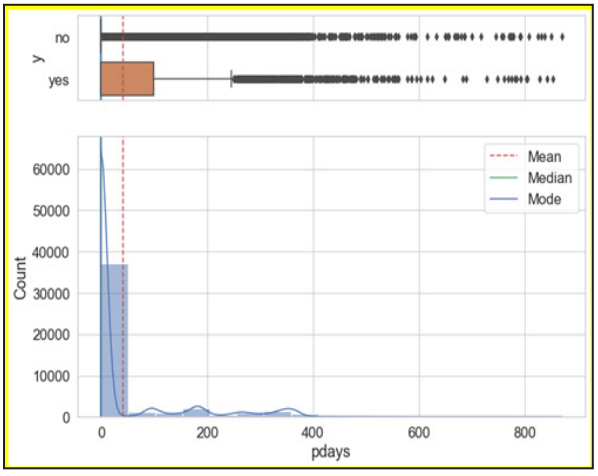

Figure 17: pdays feature analysis.

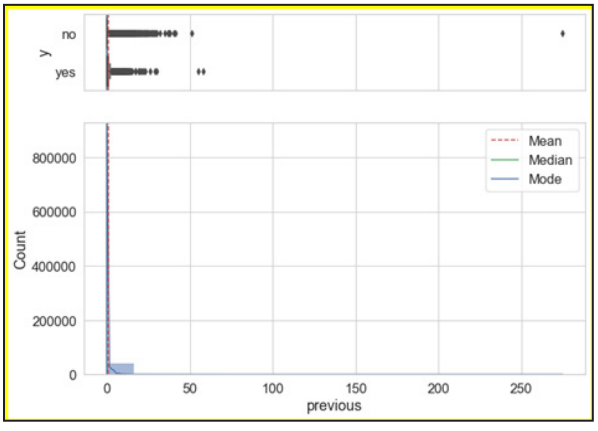

Figure 18: previous feature analysis.

3.4.Prediction

Since most machine learning algorithms operate on numerical inputs, all categorical features were transformed accordingly. We applied one-hot encoding [30] to nominal categorical variables including job, marital status, default status, housing, loan, contact type, month, and poutcome. One-hot encoding was chosen because these features have no intrinsic ordinal relationship, and this method prevents the introduction of false ordinal assumptions that might mislead the model. Additionally, one-hot encoding is well-suited for tree-based models, which can naturally handle sparse feature spaces without requiring normalization.

In contrast, the «education» feature was encoded using label encoding [31], mapping categories such as unknown = 0, primary = 1, secondary = 2, and tertiary = 3. Label encoding was chosen here to preserve the inherent ordinal nature of educational levels, which can carry meaningful ordering for certain algorithms, particularly those sensitive to rank. The target variable y was also label-encoded (no = 0, yes = 1) to prepare it for binary classification.

Alternative encoding strategies such as target encoding [32] or frequency encoding [33] were considered but not applied. While target encoding can capture the impact of categories based on the target variable's distribution, it introduces a risk of data leakage if not carefully cross-validated. Frequency encoding, though compact, may lose categorical interpretability and introduce pseudo-ordinal relationships in purely nominal data.

To optimize the performance of the Support Vector Machine (SVM) classifier, we applied random search [34], a randomized hyperparameter search method that efficiently explores a wide range of parameter combinations. Unlike grid search [34], which exhaustively evaluates all specified combinations, random search offers a more computationally efficient approach by sampling a fixed number of parameter settings from the specified distributions. This makes it particularly suitable when tuning over a large or continuous hyperparameter space.

The optimal hyperparameters were identified based on 10-fold cross-validation using the macro F1-score as the evaluation metric. Table 2 summarizes the optimal values discovered during the tuning process.

The dataset was initially divided into training and testing sets; however, due to class imbalance, adjustments were necessary. To address this, we implemented the balanced class weight technique [35] and applied 10-fold cross-validation [36] to enhance model reliability.

Table 2: Hyperparameters of SVM for Feature Selection

hyperparameter	Description	Value
C	Inverse regularization strength	0.01
loss	The loss function used for optimization	squared_ hinge
max_iter	Maximum number of optimization iterations	1000
tolerance	Stopping criterion for convergence	0.0001

To ensure robust and unbiased evaluation, a 10-fold stratified cross-validation procedure was conducted on the training dataset. For each fold, the model was trained on 90% of the data and evaluated on the remaining 10%, preserving the class distribution across splits. The performance metrics recorded for each fold include Accuracy, Precision, Recall, and F1-Score. The table below summarizes these metrics across all folds, followed by the overall average and standard deviation for each metric. This comprehensive evaluation provides insight into the model's generalization capability and consistency across different subsets of the training data, reporting and average accuracy of 90.44%. The best model is then used for testing.

Table 3: 10-fold cv results

Fold	Accuracy	Precision	Recall	F1-Score
1	0.904688	0.681481	0.347826	0.460576
2	0.902455	0.667954	0.327652	0.439644
3	0.900022	0.640000	0.332703	0.437811
4	0.905773	0.69434	0.347826	0.463476
5	0.903561	0.685259	0.325142	0.441026
6	0.904446	0.684411	0.340265	0.454545
7	0.906879	0.719512	0.334594	0.456774
8	0.903119	0.681275	0.323251	0.438462
9	0.904667	0.685606	0.342155	0.456494
10	0.904446	0.676364	0.351607	0.462687
Mean	**0.904006**	**0.68162**	**0.337302**	**0.451149**
STD	**0.001883**	**0.019948**	**0.01014**	**0.010643**

After cross-validation, the SVM classifier was trained on the complete training dataset and evaluated on the held-out test set. The confusion matrix presented below provides a detailed breakdown of the model's classification performance in terms of true positives, true negatives, false positives, and false negatives. The rows represent the actual class labels, while the columns correspond to the predicted labels. Class labels are denoted as 'Yes' (1) for the positive class and 'No' (0) for the negative class as in Figure 19.

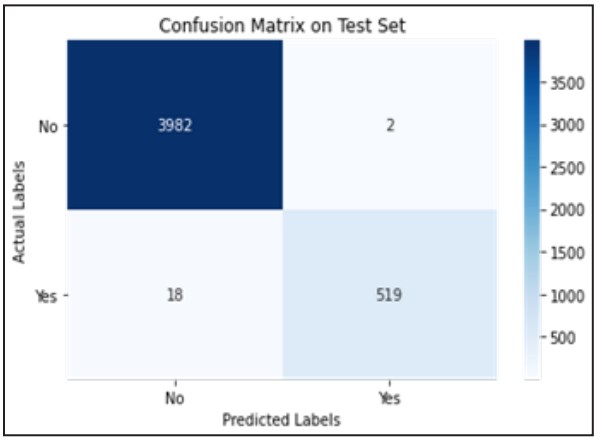

Figure 19: Confusion matrix.

This visualization enables a direct interpretation of the types of classification errors made by the model and highlights its ability to distinguish between the two classes in real-world testing scenarios. The training and testing processes were conducted using the Support Vector Machine [11] algorithm, resulting in an classifier with a 99.56% accuracy rate, 99.62% precision, 99.95% recall, and 98.11% F1-score.

To further assess the model's discriminatory power, the Receiver Operating Characteristic (ROC) curve was generated using the predicted probabilities of the positive class. The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate at various classification

thresholds. The corresponding AUC quantifies the overall ability of the model to distinguish between the positive and negative classes as in Figure 20. An AUC of 0.9819 indicates that the model demonstrates a strong capability in ranking positive instances higher than negative ones, which is especially important in imbalanced classification tasks.

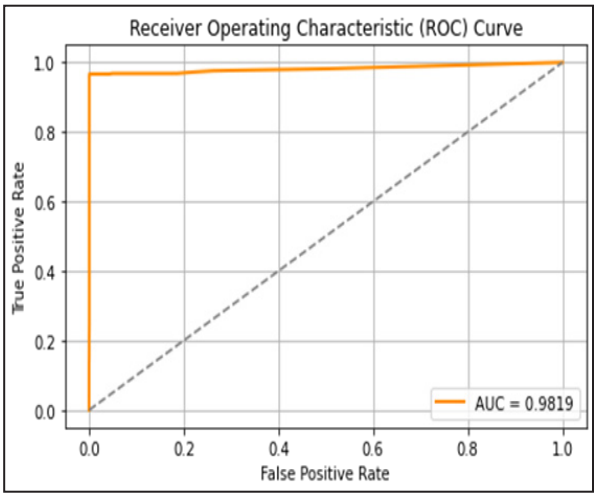

Figure 20: AUC curve.

Figure 21 presents a comparative analysis of the proposed system against existing state-of-the-art models that have utilized the bank marketing dataset.

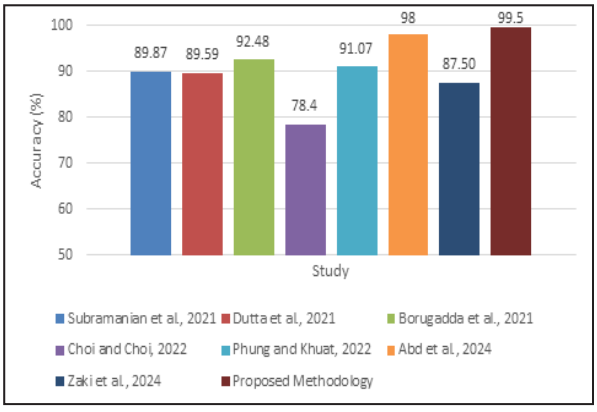

Figure 21: Comparison between systems using bank marketing dataset.

The results demonstrate that the proposed model outperforms all previous techniques, achieving superior classification accuracy. To establish a baseline for evaluating model performance, a Logistic Regression [37] classifier was trained and evaluated using the same preprocessing steps applied to the SVM model as shown in Table 4. While logistic regression offers simplicity, interpretability, and computational efficiency, it is not capturing complex, non-linear patterns inherent in telemarketing data.

In contrast, the SVM classifier significantly outperformed logistic regression across all evaluation metrics. The notable gains in precision (99.62% vs. 62.31%) and recall (99.95% vs. 30.54%) suggest that the SVM model not only identifies a higher proportion of true positives (interested clients) but also minimizes false positives, which is critical in avoiding unnecessary outreach to disinterested customers.

Table 4: Model Performance Comparison with logistic regression

Metric	Logistic Regression	SVM
Accuracy (%)	89.74	**99.56**
Precision (%)	64.31	**99.62**
Recall (%)	30.54	**99.95**
F1-Score (%)	41.41	**98.11**

While the results obtained in this study demonstrate promising performance of the classification model, several limitations should be acknowledged. First, the analysis was conducted using a specific dataset, which may not fully capture the diversity

of customer behaviors or market conditions across different regions or time periods. As a result, the generalizability of the model may be limited when applied to datasets with different distributions, feature definitions, or external economic factors. Additionally, the model›s performance is inherently dependent on the quality and completeness of the input features; any noise, bias, or imbalance in future data could affect predictive accuracy. Furthermore, although hyperparameter tuning was performed to optimize model performance, the search space was constrained due to computational limitations. Future work should consider validating the model across multiple datasets and incorporating temporal or market dynamics to enhance robustness and adaptability.

4. Conclusion

This study explored the role of predictive analytics in enhancing banking decision-making through data mining techniques. It was concluded that by leveraging machine learning models, banks can improve their ability to predict customer behavior, optimize marketing campaigns, and make more effective data-driven decisions. The research justified the added value of adopting data mining techniques, such as classification algorithms and predictive models as they enabled more efficient analysis to large volumes of banking data which in turn provided an insight into patterns and trends for more effective decision-making. The results have raised the importance of the data preprocessing tasks and the model evaluation in building more accurate predictive models. It has also ensured the necessity of handling the problem of data imbalances as well as optimizing model parameters to ensure better accuracy and robustness. Results have also demonstrated the effectiveness of applying the Support Vector Machines (SVM) and other state-of-the-art algorithms in terms of enhancing the predictive results about customer responses to financial products. To handle the class imbalance problem, we applied a balanced class weight approach along with 10-fold cross-validation. The SVM algorithm was utilized for training and testing, resulting in an optimal classifier. The proposed approach achieved 99.56% accuracy, 99.62% precision, 99.95% recall, and 98.11% F1-score and 0.9819 AUC, surpassing existing state-of-the-art techniques. Future research can focus on integrating deep learning techniques and real-time data processing to enhance predictive capabilities further. Ultimately, predictive analytics using data mining techniques presents a powerful tool for modern banking institutions, enabling them to make informed, strategic decisions that drive growth and customer satisfaction.

References

[1] U. Noreen, A. Shafique, Z. Ahmed, M. Ashfaq, Banking 4.0: Artificial Intelligence (AI) in Banking Industry & Consumer's Perspective, Sustain. (2023). https://doi.org/10.3390/su15043682.

[2] L. Wu, D. Yu, Y. Lv, Digital banking and deposit: Substitution effect of mobile applications on web

services, Financ. Res. Lett. (2023). https://doi.org/10.1016/j.frl.2023.104138.

[3] Y. Feng, Y. Yin, D. Wang, L. Dhamotharan, A dynamic ensemble selection method for bank telemarketing sales prediction, J. Bus. Res. (2022). https://doi.org/10.1016/j.jbusres.2021.09.067.

[4] B. Meyerhof Salama, V.P. Braga, The case for private administration of deposit guarantee schemes, J. Bank. Regul. (2023). https://doi.org/10.1057/s41261-021-8-00188.

[5] C. Xie, J. Le Zhang, Y. Zhu, B. Xiong, G.J. Wang, How to improve the success of bank telemarketing? Prediction and interpretability analysis based on machine learning, Comput. Ind. Eng. (2023). https://doi.org/10.1016/j.cie.2022.108874.

[6] A. Gaspar-Cunha, P. Costa, A. Delbem, F. Monaco, M.J. Ferreira, J. Covas, Evolutionary Multi-Objective Optimization of Extrusion Barrier Screws: Data Mining and Decision Making, Polymers (Basel). (2023). https://doi.org/10.3390/polym15092212.

[7] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, Electron. Mark. (2021). https://doi.org/10.1007/s125252-00475-021-.

[8] S. Tanwar, Machine Learning, in: Comput. Sci. Its Appl., 2024. https://doi.org/10.12012-9781003347484/.

[9] M.C.S. Tad, M.S. Mohamed, S.F. Samuel, M.J. Deepa, ARTIFICIAL INTELLIGENCE AND ROBOTICS AND THEIR IMPACT ON THE PERFORMANCE OF THE WORKFORCE IN THE BANKING SECTOR, Rev. Gest. Soc. e Ambient. (2023). https://doi.org/10.24857/rgsa.v17n6012-.

[10] S.R. Durugkar, R. Raja, K.K. Nagwanshi, S. Kumar, Introduction to data mining, Data Min. Mach. Learn. Appl. (2022). https://doi.org/10.10029781119792529/.ch1.

[11] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. (1998). https://doi.org/10.1023/A:1009715923555.

[12] J. Han, M. Kamber, J. Pei, Data Transformation by Normalization, 2011. https://doi.org/10.1016/B978-12-0-0-1.00001-381479.

[13] J. Asare-Frempong, M. Jayabalan, Predicting customer response to bank direct telemarketing campaign, in: 2017 Int. Conf. Eng. Technol. Technopreneurship, ICE2T 2017, 2017. https://doi.org/10.1109/ICE2T.2017.8215961.

[14] M. Moreira, C., Lima, J., & Silva, Predicting customer behavior in banking telemarketing campaigns with machine learning, Comput. Ind. Eng. 141 (2020).

[15] M. Peter, H. Mofi, S. Likoko, J. Sabas, R. Mbura, N. Mduma, Predicting customer subscription in bank telemarketing campaigns using ensemble learning models, Mach. Learn. with Appl. 19 (2025) 100618.

[16] D.Z. Abidin, M. Rosario, A. Sadikin, N. Nurhadi, J. Jasmir, Improving Term Deposit Customer Prediction Using Support Vector Machine with SMOTE and Hyperparameter Tuning in Bank Marketing Campaigns, J. Tek. Inform. 6 (2025) 1267-1278.

[17] M.F. Tanvir, M.M. Hossain, M.A. Jishan, Bayesian Regression for Predicting Subscription to Bank Term Deposits in Direct Marketing Campaigns, in: 2024 Int. Conf. Decis. Aid Sci. Appl., 2024: pp. 1-5.

[18] S. Sarlak, M., Tavafi, M., & Parsa, Predicting customer churn using machine learning models: A case study in banking, 1 (2020) 51-65.

[19] F. Safarkhani, S. Moro, Improving the accuracy of predicting bank depositor's behavior using a decision tree, Appl. Sci. (2021). https://doi.org/10.3390/app11199016.

[20] P. Borugadda, P. Nandru, C. Madhavaiah, S. Theresa, Predicting the Success of Bank Telemarketing for Selling Long-term Deposits: An Application of Machine Learning

Algorithms, St. Theresa J. Humanit. Soc. Sci. (2021).

[21] T.D. Phung, D.B. Khuat, Potential Customers Prediction in Bank Telemarketing, FPTU Ha Noi, 2022.

[22] A.M. Zaki, N. Khodadadi, W.H. Lim, S.K. Towfek, Predictive Analytics and Machine Learning in Direct Marketing for Anticipating Bank Term Deposit Subscriptions, Am. J. Bus. Oper. Res. (2024). https://doi.org/10.54216/ajbor.110110.

[23] N.S. Abd, O.S. Atiyah, M.T. Ahmed, A. Bakhit, Digital Marketing Data Classification by Using Machine Learning Algorithms., Iraqi J. Electr. \& Electron. Eng. 20 (2024).

[24] S. Moro, P. Cortez, P. Rita, A data-driven approach to predict the success of bank telemarketing, Decis. Support Syst. (2014). https://doi.org/10.1016/j.dss.2014.03.001.

[25] T. Kim, W. Ko, J. Kim, Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting, Appl. Sci. (2019). https://doi.org/10.3390/app9010204.

[26] P. Kolte, M. Wolfe, Elimination of Redundant Array Subscript Range Checks, ACM SIGPLAN Not. (1995). https://doi.org/10.1145223428.207160/.

[27] A.G. Dufera, T. Liu, J. Xu, Regression models of Pearson correlation coefficient, Stat. Theory Relat. Fields. (2023). https://doi.org/10.108024754269.2023.21/64970.

[28] T.L. Weissgerber, N.M. Milic, S.J. Winham, V.D. Garovic, Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm, PLoS Biol. (2015). https://doi.org/10.1371/journal.pbio.1002128.

[29] S.R.D. Putranti, Relationships between Mean, Median and Mode, Aloha Int. J. Multidiscip. Adv. (2021). https://doi.org/10.33846/aijmu30403.

[30] S. Bagui, D. Nandi, S. Bagui, R.J. White, Machine Learning and Deep Learning for Phishing Email Classification using One-Hot Encoding, J. Comput. Sci. (2021). https://doi.org/10.3844/jcssp.2021.610.623.

[31] C. Herdian, A. Kamila, I.G. Agung Musa Budidarma, Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi, Technol. J. Ilm. (2024). https://doi.org/10.31602/tji.v15i1.13457.

[32] F. Pargent, F. Pfisterer, J. Thomas, B. Bischl, Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features, Comput. Stat. (2022). https://doi.org/10.1007/s001806-01207-022-.

[33] Y. Chen, H. Li, Stochastic Computing Using Amplitude and Frequency Encoding, IEEE Trans. Very Large Scale Integr. Syst. (2022). https://doi.org/10.1109/TVLSI.2022.3150569.

[34] D.A. Anggoro, N.A. Afdallah, Grid Search CV Implementation in Random Forest Algorithm to Improve Accuracy of Breast Cancer Data, Int. J. Adv. Sci. Eng. Inf. Technol. (2022). https://doi.org/10.18517/ijaseit.12.2.15487.

[35] G. Sambasivam, G.D. Opiyo, A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks, Egypt. Informatics J. (2021). https://doi.org/10.1016/j.eij.2020.02.007.

[36] S.M. Malakouti, M.B. Menhaj, A.A. Suratgar, The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction, Clean. Eng. Technol. (2023). https://doi.org/10.1016/j.clet.2023.100664.

[37] L. Connelly, Logistic regression, MEDSURG Nurs. (2020). https://doi.org/10.466929781847423399.014/.