Journal of Theoretical and Applied Information Technology
15th October 2024. Vol.102. No. 19
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING BANKING SERVICES THROUGH SMART DATA ANALYTICS FRAMEWORK

YASMINE E. EL-GEMEIE [1], MOHAMED ABDELSALAM [2], IBRAHIM F. MOAWAD [3]

[1, 2] Information Systems Department, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt
[3] Professor of Artificial Intelligence, Faculty of Computer and Information sciences, Ain Shams University, Cairo, Egypt
Faculty of Computer Science and Engineering, New Mansoura University, Dakhlia, Egypt
E-mail:[1]yasmine.emad-el-din21@commerce.helwan.edu.eg, [2]dr.m_abdelsalam@commerce.helwan.edu.eg, [3]ibrahim_moawad@cis.asu.edu.eg

ID 55432 Submission	Editorial Screening	Conditional Acceptance	Final Revision Acceptance
25-08-2024	26-09-2024	07-09-2024	01-10-2024

ABSTRACT

Banking targeted marketing strategies have undergone evolution with the integration of predictive analytics and machine learning techniques which play a pivotal role in engaging customers and enticing them to subscribe to various packages and fixed-term deposit offers. The core problem identified was imprecise customer segmentation, resulting in less accurate predictions. The present study focuses on increasing banking sales by predicting customer reactions accurately, contributing to personalized interactions, nurturing customer relationships and proposes the development of a prediction model using machine learning algorithms offering predictive capabilities for sales, customer preferences, new client identification, and efficiency gains. The methodology encompasses data exploration, visualization, preprocessing techniques are applied and implementation of various machine learning models, including XGBoost, Random Forest, Decision Tree, KNN, Logistic Regression, and Naive Bayes. Using a large dataset from a Portuguese bank from Kaggle are employed which used for detection of customer reactions to fixed-term deposit subscriptions. The present results demonstrate high accuracy rate of 93.48% using Random Forest and 92.06% using XGBoost compared to other studies. The consistency observed in cross-validation suggests the models' robustness, emphasizing their suitability for real-world banking campaigns, enhance customer segmentation, optimize targeting, recommend suitable products and improve overall efficiency. While the results are promising, future work should focus on hyperparameter optimization and further refinement of ensemble techniques to boost predictive accuracy.

Keywords: *Artificial intelligence, Machine learning, Banking services, Bank marketing, Decision tree, K-nearest neighbors' algorithm, Naive Bayes, Support vector machines.*

1. INTRODUCTION

In the highly competitive and fast-evolving banking sector, staying ahead requires more than just traditional strategies; it demands the adoption of cutting-edge technologies to meet the growing expectations of customers and improve operational efficiency. One critical challenge faced by banks is the ability to accurately predict customer behavior, particularly in the context of product subscriptions, such as term deposits. Inaccurate customer segmentation and unreliable predictions can lead to missed opportunities, inefficient resource allocation, and diminished customer satisfaction. As banks continue to shift toward data-driven decision-making, machine learning (ML) and artificial intelligence (AI) have emerged as transformative tools capable of addressing these challenges [1].

The significance of improving customer segmentation and predicting customer behavior in banking cannot be overstated. Effective

customer targeting not only enhances marketing efforts but also strengthens customer relationships, drives revenue growth, and ultimately contributes to the bank's competitive advantage. In the realm of direct marketing campaigns, where banks reach out to clients for term deposit subscriptions, the need for precise and reliable predictions becomes even more crucial. The banking industry has historically relied on traditional methods that often yield suboptimal results, leading to wasted marketing efforts and resources [1]–[3].

This study addresses the core issue of improving customer segmentation and predictive accuracy in banking services by exploring the potential of machine learning algorithms to enhance targeted marketing efforts. The primary problem lies in the inability of sales agents to accurately identify which customers are likely to subscribe to a term deposit, leading to inefficient campaigns and reduced customer engagement. This issue is critical for banks, as inaccurate predictions hinder growth, affect customer trust, and increase operational costs [4]-[5].

By leveraging machine learning algorithms such as decision trees, random forests, and support vector machines. This research aims to build a predictive framework that can revolutionize how banks approach customer targeting and marketing. Using the Bank Marketing Data Set from the UCI Machine Learning Repository, the study will explore how these algorithms can analyze historical customer interaction data to forecast future behavior, providing actionable insights into customer preferences. This investigation will offer a data-driven approach to improving customer segmentation, helping banks to deploy resources more effectively, engage customers more successfully, and boost overall performance in a highly competitive market[6]–[8].

This paper not only contributes to the growing body of research on AI and ML in the banking sector but also aims to provide practical solutions for overcoming the challenges of inaccurate predictions in real-world marketing scenarios. Through a comprehensive exploration of various ML models and their application to direct marketing data, we aim to demonstrate how banks can harness the power of AI to enhance customer engagement, streamline operations, and thrive in a rapidly changing industry [9]– [12].

2. LITERATURE REVIEW

The increasing number of Internet users and the rapid proliferation of online businesses have led to a surge in interconnected e-commerce applications, with online banking emerging as one of the most prominent sectors since the 1980s. The digitization of banking services has not only streamlined financial transactions but has also introduced new challenges in terms of customer targeting and marketing efficiency. To tackle these challenges, researchers have turned to advanced data-driven techniques, such as machine learning (ML) and artificial intelligence (AI), to enhance predictive accuracy and operational efficiency in the banking sector.

One study on financial transactions demonstrated that Support Vector Machine (SVM) outperformed traditional logistic regression models in predicting short-term outcomes, highlighting its potential for improving the intelligence and security of online banking systems [13], [14], [15]. This discovery has encouraged further exploration of more sophisticated machine learning models for enhancing banking services. Financial services domains have widely applied techniques like deep learning, data mining, decision trees, and k-nearest neighbors (k-NN), demonstrating promising results in predicting customer behaviors and improving service delivery [16]–[18].

Zaki et al. (2024) investigated how ML-based predictive analytics has transformed direct marketing in the banking industry. Using the Kaggle dataset, the study applied multiple models, including the SGD classifier, k-NN classifier, and random forest, to forecast customer subscriptions to term deposits. The Random Forest Classifier achieved an accuracy of 87.5%, with strong metrics in both positive predictive value (PPV) and negative predictive value (NPV), demonstrating its utility for banks aiming to refine their marketing strategies in a competitive financial landscape [13].

The rise of machine learning in banking has also introduced ethical and legal concerns, especially in the context of biased algorithms. Radovanović et al. (2021) examined the potential biases in prediction models, particularly regarding race and gender, which could lead to unfair treatment of customers. In their study, they proposed a fair classifier chain machine learning model to tackle multi-label

classification issues and minimize bias in bank marketing applications. While fairness improved by 7% to 17%, the accuracy of predictions dropped by up to 9% in terms of AUC (Area Under Curve), indicating a critical trade-off between fairness and precision [18].

Feature selection is another key area in optimizing the performance of machine learning models for banking applications. Roy et al. (2021) explored the use of genetic algorithms to improve the accuracy and efficiency of ML algorithms. By selecting relevant features from datasets such as the Portuguese bank dataset and Yahoo Apple Inc. stock data, the genetic algorithm reduced the number of features required by up to 50% and enhanced prediction accuracy by 10%, making it a valuable tool for financial forecasting [19].

The Portuguese bank dataset has been a focal point for many studies, providing a rich dataset for analyzing customer behavior. Duwairi & Halloush (2022) tested various ML models, including Logistic Regression, Naive Bayes, and Random Forest, with Random Forest achieving an impressive accuracy of 95.2%. Similar studies on other datasets, such as stock market data and gold prices, further validated the effectiveness of ML models like linear regression and artificial neural networks (ANNs) in predicting financial outcomes [20].

Hayder et al. (2023) explored the use of machine learning in targeted advertising for fixed-term deposit offers. The study applied decision trees, SVM, and other classifiers to predict customer responses to marketing campaigns, achieving an accuracy of 91% with decision trees and 89% with SVM. This research underscores the potential of ML models to enhance the precision of customer targeting, thus boosting the effectiveness of banking advertisements and increasing conversion rates [21].

Although previous studies explored the utilization of predictive analytics and machine learning models to foresee bank term deposit subscriptions, but still further studies is required to enhance the effectiveness of these technologies to refinement the precision of direct marketing strategies in the banking sector with quantification and comparison of their performance using key metrics such as accuracy. Also further studies is required to declare the role of client demographics, financial history, and interactions with marketing campaigns in anticipation of bank term deposit subscriptions which we tried to declare in the present study.

3. METHOD

The methodology employed for predicting customer reactions in the context of direct marketing campaigns involves a systematic four-step process to process the data effectively. The initial step focuses on eliminating instances with unknown data, thereby enhancing the dataset's reliability. The subsequent step revolves around the conversion of categorical features into numeric representations to facilitate machine learning algorithms' application. Following this, the data undergoes a balancing process to ensure fairness in model training. The final step involves the selection of optimal features for the predictive model [22]- [38].

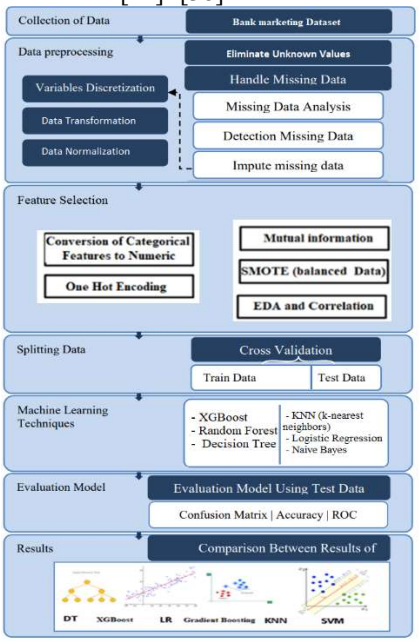

Figure 1: The Proposed Model

3.1 DATA Set

For our analysis, we use real data from a Portuguese bank that comes from advertising efforts to sell long-term savings that took place over five years, from May 2008 to June 2013. All conversations are made over the phone with a real person acting as the go-between. Most of the time the few incoming contacts are used when a client calls the bank for some other reason, and the agent uses the call to try to sell the deposit. Most of the contacts are outgoing.

The dataset has 52944 contacts made over the phone, but only 6557 of them led to successful deposit

orders. This means that the dataset is not fair. The previous study [23] looked at 150 features that described the people, but after careful selection, they only used 22 features for projected. This group will be our starting point, and we will add more related features to it.

The dataset pertains to direct marketing campaigns conducted by a Portuguese banking institution, primarily based on phone calls. The objective was to determine whether clients would subscribe ('yes') or not ('no') to the offered product, a bank term deposit [23].

3.2 Pre-processing

The data cleaning process commenced with the exclusion of features containing 330 unknown attributes, addressing the challenge of null values. Categorical data were transformed into numeric formats, and exploratory data analysis (EDA) was conducted to discern underlying patterns in the dataset [24].

As part of the preprocessing step, the information is made ready for Python code to handle and import. It's possible that adjusting the missing values or turning the data into numbers is part of the planning. Values had to be turned into numbers in this dataset so that it could be imported and tested by different methods and data normalization is applied to the data.

1. Handling Unknown Data: The initial step involves the identification and removal of instances with unknown or missing data. This ensures a clean and consistent dataset for analysis, preventing potential distortions in the results due to incomplete information. The major benefit of getting rid of cases that are missing data is that it makes sure you have a clean sample to analyze. Using incomplete data incorrectly can change results and lead to wrong findings.

2. Conversion of Categorical Features to Numeric: Categorical features within the dataset are converted to numeric representations to facilitate compatibility with machine learning algorithms. This step is crucial for ensuring uniform data types and enabling the effective utilization of mathematical models for analysis. **Allows mathematical operations**: This is the main advantage. When category data is turned into numbers, it can be used in mathematical processes. This is important for machine learning methods that use computing distances, means, and other math numbers as part of the learning process.

3. Data Balancing: Given the inherent imbalances that might exist in the dataset, particularly in the context of banking success, the third step focuses on data balancing. Techniques such as Synthetic Minority Over-sampling Technique (SMOTE) are employed to address class imbalances. This ensures that the predictive model is not skewed toward the majority class, allowing for more accurate insights into the success factors.

4. Feature Selection using Mutual Information: To enhance the efficiency of the model and focus on the most influential factors, a feature selection process is implemented. Mutual Information, a measure of the dependency between variables, is utilized to identify and retain the most informative features. This step contributes to a streamlined dataset, optimizing computational resources while maintaining the integrity of critical information.

   Mutual information checks how dependent variables are on each other to find the most useful traits that are connected to the main variable. It helps pick qualities that can predict output.

   The utilization of mutual information simplifies datasets by selecting the most informative attributes. Getting rid of traits that aren't needed simplifies and computationally improves the model.

3.3 Classification algorithms

The proposed model for accurate prediction integrates various machine learning classifiers and feature selection techniques. The model comprises three key steps: feature selection, data pre-processing, and the application of diverse classifiers. During the pre-processing phase, missing values are addressed, and data normalization techniques are applied. The selection of crucial features involves utilizing methods such as mutual information, chi-squared, and Pearson correlation. Subsequently, classifiers like Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Logistic Regression (LR) are employed on the output of the feature selection step. The overall model's effectiveness is evaluated using a set of metrics during the results interpretation phase. Figure 1 illustrates the comprehensive

framework of the proposed model, and the subsequent section will provide detailed insights into each stage of this methodology [25] – [30].

The study employs a diverse set of machine learning classifiers to predict customer reactions in the realm of direct marketing campaigns. These classifiers include Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Logistic Regression (LR). Each classifier brings a unique approach to the predictive modeling process [30] – [36].

- Random Forest (RF): RF is an ensemble learning method that constructs a multitude of decision trees during training and merges them to enhance predictive accuracy and control overfitting [30] – [36].
- Decision Tree (DT): DT is a tree-like model where internal nodes represent decisions based on input features, and leaf nodes represent the predicted outcomes [30] – [36].
- Support Vector Machine (SVM): SVM is a supervised learning algorithm that classifies data by finding the hyperplane that best separates different classes in the feature space [30] – [36].
- K-Nearest Neighbors (KNN): KNN is a non-parametric, instance-based learning algorithm that classifies new data points based on the majority class of their k-nearest neighbors in the feature space [30] – [36].
- Logistic Regression (LR): LR is a linear model used for binary classification, predicting the probability of an instance belonging to a particular class [30] – [36].

By leveraging the strengths of these diverse classifiers, the study aims to comprehensively evaluate and compare their performance in predicting customer reactions based on transaction data. Each method brings its unique characteristics, and their collective application allows for a robust analysis of the predictive model's effectiveness.

4 RESULTS AND DISCUSSION

The following section shows the results of the trial and the rating measures. These things can be found on the PC that was used for the experiments: Windows 10 and an Intel(R) Core(TM) i5-9750H x64-based CPU with 2.60 GHz and 2.59 GHz speeds. Eight gigabytes of RAM. The other 70% of the data is used for training and validation, while the other 30% was used for testing. The model does its job with the help of Python code.

Exploratory Data Analysis (EDA) is an important part of the data analysis process. It involves looking at a dataset both visually and statistically to find trends, gain useful insights, and find possible connections between factors.

Table 1: EDA

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
count	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.024060	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911	0.112654
std	10.421250	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528	0.316173
min	17.000000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000	0.000000
25%	32.000000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000	0.000000
50%	38.000000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000	0.000000
75%	47.000000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000	0.000000
max	98.000000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000	1.000000

Analyzing all the data is the goal of the class. The class will figure out what kind of data the column is and show it in the right way. It will spread out the data in cells whose data type is integer and show us the mean, standard deviation, and minimum numbers on its own. When we have classified data, we can see how much of it comes from each group.

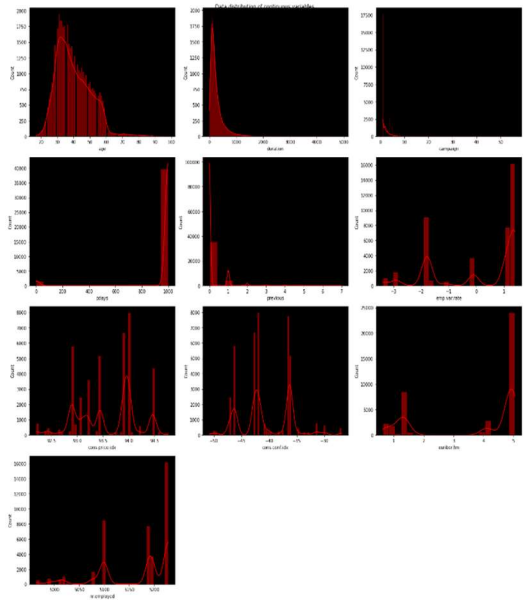

Figure 2: Data distribution of continuous variables.

A correlation matrix is a type of statistics that can be used to look at the link between two factors in a set of data. Figure 3 shows how to use this method. A correlation coefficient of -1 means that the relationship between two factors is weak, a correlation value of 1 means that the relationship is strong, and a correlation coefficient of 0 means that the relationship is neutral. There are association values in each cell of the matrix, which is a table.

Journal of Theoretical and Applied Information Technology
15th October 2024. Vol.102. No. 19
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

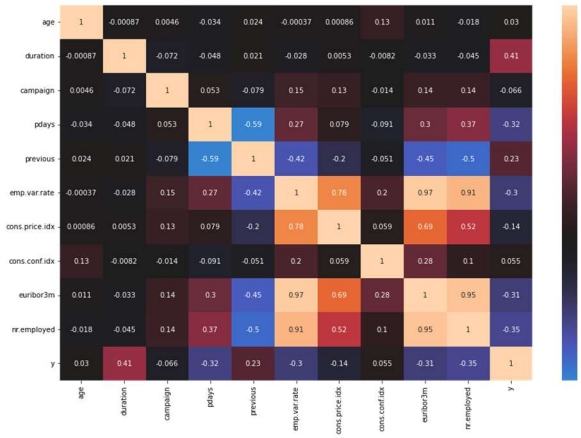

Figure 3: the correlation matrix.

Figure 4 shows the pattern or spread of numbers that the variable takes on is called its data distribution. The goal variable is the variable we want to guess or understand from other data in the dataset. This is used in machine learning and statistical analysis. It is very important to understand how the target variable's data is distributed because it gives us information about the variable we are trying to model.

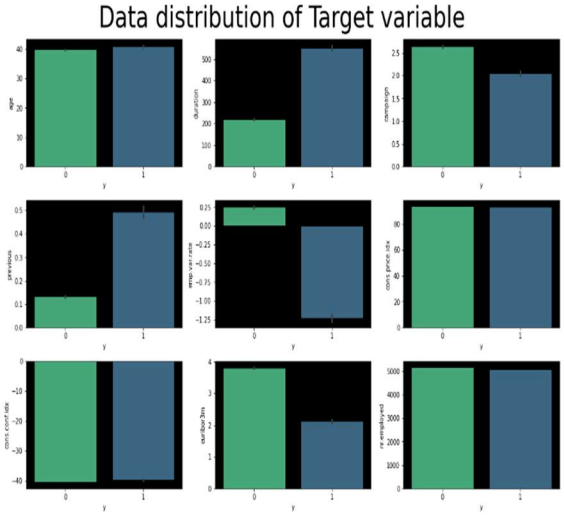

Figure 4: Data distribution of Target variable

We won't get very good results when we use the dataset with the algorithms. But we need to process the information first before we can use algorithms. There was a big difference between the balanced dataset and the unbalanced dataset because the unbalanced dataset did not do well.

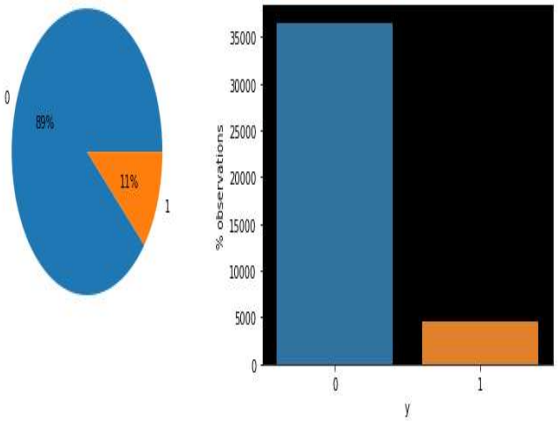

Figure 5: unbalanced dataset

To make balanced datasets using SMOTE is a popular method that generates synthetic instances of the minority class, addressing the imbalance by creating new, realistic data points and Ensemble methods like Random Forest and Gradient Boosting can be effective, as they inherently handle imbalanced datasets by combining predictions from multiple models.

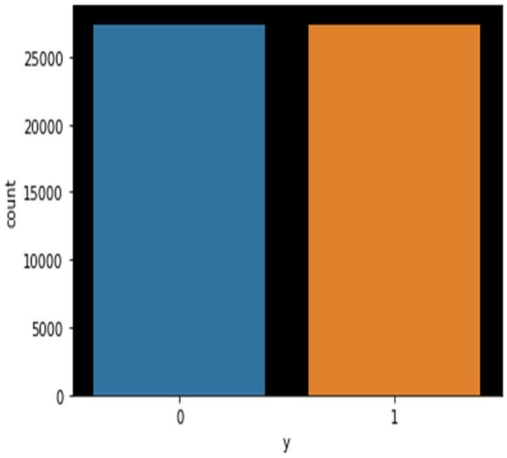

Figure 6: balanced dataset

A method called "One-Hot Encoding" is used to prepare data, especially when working with category factors. You can use it to turn category data, like names or text, into a number format that machine learning systems can understand. For this process, two binary columns are made for each name or category. For each data point, only one of these columns is marked as "1" (hot), which means it belongs to that category.

Table 2: One-Hot Encoding

	age	duration	campaign	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	job_blue-collar	job_entrepreneur	job_housemaid
0	56	261	1	0	1.10	93.99	-36.40	4.86	5,191.00	0	0	1
1	57	149	1	0	1.10	93.99	-36.40	4.86	5,191.00	0	0	0
2	37	226	1	0	1.10	93.99	-36.40	4.86	5,191.00	0	0	0
3	40	151	1	0	1.10	93.99	-36.40	4.86	5,191.00	0	0	0
4	56	307	1	0	1.10	93.99	-36.40	4.86	5,191.00	0	0	0

4.1 Classification models

4.1.1 Logistic Regression

Table 3: Evaluation matrix for Logistic Regression

Model Accuracy	85.30%			
Model F1-Scoe	83.38%			
Cross Val Accuracy	85.50 %			
Cross Val Standard Deviation	0.55 %			
	Precision	**recall**	**f1-score**	**support**
0	0.86	0.98	0.91	8013
1	0.83	0.42	0.56	2284
Accuracy			0.85	10297
Macro average	0.84	0.70	0.74	10297
Weighted average	0.85	0.85	0.83	10297

Table 3 shows that the predicted values by using Logistic Regression algorithm are as follows: Model Accuracy: 85.21%, Model F1-Score: 83.28% and Cross Val Accuracy: 85.40 %.

Model Accuracy	76.84%			
Model F1-Score	72.98%			
Cross Val Accuracy	78.67 %			
Cross Val Standard Deviation	0.17 %			
	Precision	**recall**	**f1-score**	**support**
0	0.77	0.97	0.85	7256
1	0.78	0.30	0.43	3041
accuracy			0.77	10297
Macro average	0.77	0.63	0.64	10297
weighted average	0.77	0.77	0.73	10297

4.1.2 Decision Tree

Table 4: Evaluation matrix for Decision Tree

Model Accuracy	88.88%			
Model F1-Score	88.69%			
Cross Val Accuracy	91.30 %			
Cross Val Standard Deviation	14.66 %			
	precision	**recall**	**f1-score**	**support**
0	0.93	0.94	0.94	9044
1	0.55	0.51	0.53	1253
Accuracy			0.89	10297
Macro average	0.74	0.72	0.73	10297
Weighted average	0.89	0.89	0.89	10297

Table 4 shows that the predicted values by using Decision Tree algorithm are as follows: Model Accuracy: 88.88%, Model F1-Score: 88.69% and Cross Val Accuracy: 91.30 %.

4.1.3 K-nearest neighbor

Model Accuracy	87.52%			
Model F1-Score	86.51%			
Cross Val Accuracy	91.88 %			
Cross Val Standard Deviation	1.43 %			
	Precision	**Re-call**	**f1-score**	**support**
0	0.90	0.96	0.93	8580
1	0.69	0.46	0.55	1717
Accuracy			0.88	10297
Macro average	0.79	0.71	0.74	10297
Weighted average	0.86	0.88	0.87	10297

Table 5: Evaluation matrix for K-nearest neighbor

Table 5 shows that the predicted values by using **K-nearest neighbor** algorithm are as follows: Model Accuracy : 87.52% , Model F1-Score : 86.51% and Cross Val Accuracy: 91.88 %.

4.1.4 Gaussian Naive Bayes (GaussianNB)

Table 6: Evaluation matrix for GaussianNB

Table 6 shows that the predicted values by using **GaussianNB** algorithm are as follows: Model Accuracy: 76.84%, Model F1-Score: 72.98% and Cross Val Accuracy: 78.67 %.

4.1.5 Random Forest

Table 7: Evaluation matrix for Random Forest

Model Accuracy		90.44%		
Model F1-Score		90.54%		
Cross Val Accuracy		93.48 %		
Cross Val Standard Deviation		6.15 %		
	precision	recall	f1-score	support
0	0.95	0.94	0.95	9187
1	0.55	0.58	0.57	1110
Accuracy			0.90	10297
Macro average	0.75	0.76	0.76	10297
Weighted average	0.91	0.90	0.91	10297

Table 7 shows that the predicted values by using **Random Forest** algorithm are as follows: Model Accuracy: 90.44%, Model F1-Score: 90.54% and Cross Val Accuracy: 93.54 %.

4.1.6 Extreme Gradient Boosting (XGB)

Table 8: Evaluation matrix for extreme Gradient Boosting (XGB)

Model Accuracy		90.89%		
Model F1-Score		90.94%		
Cross Val Accuracy		92.06 %		
Cross Val Standard Deviation		9.31 %		
	Precision	recall	f1-score	support
0	0.95	0.95	0.95	9167
1	0.58	0.60	0.59	1130
Accuracy			0.91	10297
Macro average	0.77	0.77	0.77	10297
Weighted average	0.91	0.91	0.91	10297

Table 8 shows that the predicted values by using **XGB** algorithm are as follows: Model Accuracy: 90.89%, Model F1-Score: 90.94% and Cross Val Accuracy: 92.31%.

4 Discussion

The results of our work showcase the performance metrics of various machine learning models applied to the task of predicting customer reactions in the context of a transaction-based approach. The following table summarizes the key metrics, including model accuracy, F1-score, cross-validation accuracy, and cross-validation standard deviation.

Our approach includes multiple models such as XGBoost, Random Forest, Decision Tree, KNN, Logistic Regression, and Naive Bayes, as well as a larger dataset and framework, resulted in superior overall accuracy compared to the study in [37].

In [37], the highest accuracy achieved using a single model (Decision Tree) was 89.81% with an error rate of 10%. In contrast, the maximum accuracy obtained in our study was 93.48% using Random Forest with a CV standard deviation of only 6.15%, and XGBoost achieved an accuracy of 92.06% and CV STD of 9.31%.

Models like Naive Bayes produced comparatively lower accuracy in our case similar to [37], despite our implementation attaining 76.84% accuracy versus 84.93% in [37]. This suggests the inherent weakness of Naive Bayes assumptions for complex prediction tasks. This comprehensive compare show the differences between our results compared to other literatures. It explores innovative methodologies and approaches to propel the banking sector toward enhanced decision-making and strategic planning.

Table 9:models_comparison

Model	Acc-uracy	F1-Score	CV Accuracy	CV std
Xg Boost	90.89%	90.94 %	92.06 %	9.31%
Random Forest	90.44%	90.54 %	93.48 %	6.15%
Decision Tree	87.75%	87.41 %	91.01 %	14.23 %
KNN	87.52%	86.51 %	91.88 %	1.43%
Logistic Regression	85.30%	83.38 %	85.50 %	0.55%
Naive Bayes	76.84%	72.98 %	78.67 %	0.17%

Table 10: Maximum Score in each Column

Model	Accuracy	F1-Score	CV Accuracy	CV std
Xg Boost	90.89%	90.94 %	92.06%	9.31%
Random Forest	90.44%	90.54 %	93.48%	6.15%
Decision Tree	87.75%	87.41 %	91.01%	14.23 %
KNN	87.52%	86.51 %	91.88%	1.43%
Logistic Regression	85.30%	83.38 %	85.50%	0.55%
Naive Bayes	76.84%	72.98 %	78.67%	0.17%

Table 11: Minimum Score in each Column

Model	Accuracy	F1-Score	CV Accuracy	CV std
Xg Boost	90.89%	90.94 %	92.06%	9.31%
Random Forest	90.44%	90.54 %	93.48%	6.15%
Decision Tree	87.75%	87.41 %	91.01%	14.23 %
KNN	87.52%	86.51 %	91.88%	1.43%
Logistic Regression	85.30%	83.38 %	85.50%	0.55%
Naive Bayes	76.84%	72.98 %	78.67%	0.17%

Table 12: source [37] results

Models	Score
Gradient Boosting	0.914306
XGBoost	0.913584
Random Forest Classifier	0.910178
Logistic Model	0.909726
K-Near Neighbors	0.904815
Decision Tree Classifier	0.883693
Support Vector Machine	0.855640
Gausian NB	0.844432

The dataset analyzed in Decoding Customer Behavior: A Data-Driven Marketing Campaign Analysis pertains to direct marketing campaigns conducted by a Portuguese banking institution via phone calls. The objective was to determine whether clients would subscribe ('yes') or not ('no') to the offered product, a bank term deposit.

Our approach use a larger dataset from the same UCI bank marketing dataset, resulted in superior overall accuracy compared to the single model approaches evaluated in prior studies.

In [37], the highest accuracy achieved using a single model (Decision Tree) was 89.81% with an error rate of 10% on this dataset. In contrast, the maximum accuracy obtained in our study using the full bank marketing dataset was 93.48% using Random Forest with a CV standard deviation of only 6.15%, and XGBoost achieved an accuracy of 92.06% and CV STD of 9.31%.

The current study shares similarities with recent research in using real-world datasets and machine learning techniques for imbalanced class distributions. However, it differs in scope and methodology, utilizing a larger dataset (52,944 contacts) from a Portuguese bank spanning 2008-2013, focused on phone-based marketing campaigns. The study's initial feature set of 22 variables, with plans for expansion, contrasts with more limited variable sets in other studies, potentially allowing for more comprehensive analysis. A key difference lies in the approach to data preprocessing and model selection. Unlike studies emphasizing extensive preprocessing techniques, the current methodology's preprocessing steps are less explicitly defined. This presents an opportunity to enhance the approach with more robust data handling techniques, particularly for missing values. While other studies specify models like XGBoost and Decision Trees, the current model selection process is still under consideration, allowing for tailored algorithm selection specific to term deposit subscription prediction. Future work will focus on detailing preprocessing steps, addressing data quality issues, and carefully selecting and justifying modeling techniques to strengthen the research's robustness and comparability in financial sector predictive analytics.

By applying various machine learning algorithms to comprehensively analyze customer attributes and campaign response patterns in this dataset, the bank will be able to enhance customer segmentation, optimize targeting of prospective subscribers, recommend suitable products, reduce human error and achieve greater efficiency - ultimately translating to an improved customer experience.

While our models performed well overall, optimizing hyperparameters and evaluating

ensemble techniques could further boost predictive power when applied in real banking campaigns.

5 CONCLUSION

This paper highlights the transformative potential of machine learning in predicting customer behavior based on transaction data, providing critical insights for the banking industry. By evaluating multiple machine learning algorithms, including XGBoost, Random Forest, Decision Tree, KNN, Logistic Regression, and Naive Bayes, the study demonstrates the ability of these models to enhance personalized and effective customer interactions. Most models achieved over 85% accuracy, with Random Forest and XGBoost reaching accuracies of 93.48% and 92.06%, respectively. The low cross-validation standard deviations emphasize the robustness of these models when applied to unseen data, making them well-suited for real-world banking campaigns.

The current study shares similarities with recent research in using real-world datasets and machine learning techniques for imbalanced class distributions. However, it differs in scope and methodology, utilizing a larger dataset (52,944 contacts) from a Portuguese bank spanning 2008-2013, focused on phone-based marketing campaigns. The study's initial feature set of 22 variables, with plans for expansion, contrasts with more limited variable sets in other studies, potentially allowing for more comprehensive analysis. A key difference lies in the approach to data preprocessing and model selection. Unlike studies emphasizing extensive preprocessing techniques, the current methodology's preprocessing steps are less explicitly defined. This presents an opportunity to enhance the approach with more robust data handling techniques, particularly for missing values. While other studies specify models like XGBoost and Decision Trees, the current model selection process is still under consideration, allowing for tailored algorithm selection specific to term deposit subscription prediction. Future work will focus on detailing preprocessing steps, addressing data quality issues, and carefully selecting and justifying modeling techniques to strengthen the research's robustness and comparability in financial sector predictive analytics.

We also recommend collaborating with the industry to assess the generalizability of the models and their potential to improve customer experiences.

REFERENCES

[1] Jáuregui-Velarde, R., Andrade-Arenas, L., Molina-Velarde, P., & Yactayo-Arias, C. (2024, February 1). Financial revolution: a systemic analysis of artificial intelligence and machine learning in the banking sector. *International Journal of Electrical and Computer Engineering* (IJECE), 14(1),1079. https://doi.org/10.11591/ijece.v14i1.pp1079-1090.

[2] H. Haddad, "The effect of artificial intelligence on the AIS excellence in Jordanian Banks," *Montenegrin Journal of Economics*,vol. 17, no. 4, pp. 155–166, Sep. 2021, doi: 10.14254/1800-5845/2021.17-4.14.

[3] M. Nazar, M. M. Alam, E. Yafi, and M. M. Su'ud, "A systematic review of human-computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques," *IEEE Access*, vol. 9, pp. 153316–153348, 2021, doi:10.1109/ACCESS.2021.3127881.

[4] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695,Sep. 2021, doi: 10.1007/s12525-021-00475-2.

[5] J. R. Asor, J. L. Lerios, S. B. Sapin, J. O. Padallan, and C. A. C. Buama, "Fire incidents visualization and pattern recognition using machine learning algorithms," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 22, no. 3, pp. 1427–1435, Jun. 2021, doi: 10.11591/ijeecs.v22.i3.pp1427-1435.

[6] A. Tammenga, "The application of artificial intelligence in banks in the context of the three lines of defence model," *Maandblad Voor Accountancy en Bedrijfseconomie*, vol. 94, no. 5/6, pp. 219–230, Jun. 2020, doi: 10.5117/mab.94.47158.

[7] I. M. Enholm, E. Papagiannidis, P. Mikalef, and J. Krogstie, "Artificial intelligence and business value: a literature review," *Information Systems Frontiers*, vol. 24, no. 5, pp. 1709–1734, Oct. 2022, doi: 10.1007/s10796-021-10186-w.

[8] I. H. Sarker, "Machine learning: algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, May 2021, doi: 10.1007/s42979-021-00592-x.

[9] G. Luo, W. Li, and Y. Peng, "Overview of intelligent online banking system based on HERCULES architecture," *IEEE Access,* vol.

8, pp. 107685–107699,2020,doi: 10.1109/ACCESS.2020.2997079.

[10] H. Razavi, H. Sarabadani, A. Karimisefat, and J.-F. LEBRATY, "Profitability prediction for ATM transactions using artificial neural networks: A data-driven analysis," *in 2019 5th Conference on Knowledge Based Engineering and Innovation* (KBEI), Feb. 2019, pp. 661–665, doi: 10.1109/KBEI.2019.8735037.

[11] M. Dahmani and M. Guerti, "Vocal folds pathologies classification using Naïve Bayes networks," *in 2017 6th International Conference on Systems and Control* (ICSC), May 2017, pp. 426–432, doi: 10.1109/ICoSC.2017.7958686.

[12] Y. Benmahamed, Y. Kemari, M. Teguar, and A. Boubakeur, "Diagnosis of power transformer oil using KNN and Naive Bayes classifiers," *in 2018 IEEE 2nd International Conference on Dielectrics* (ICD),Jul.2018,pp.1–4,doi: 10.1109/ICD.2018.8468532.

[13] 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. https://doi.org/10.54216/ajbor.110110

[14] K. R. Singh, K. P. Neethu, K. Madhurekaa, A. Harita, and P. Mohan, "Parallel SVM model for forest fire prediction," Soft Computing Letters, vol. 3, Dec. 2021, doi: 10.1016/j.socl.2021.100014.

[15] L.-L. Li, Z.-F. Liu, M.-L. Tseng, K. Jantarakolica, and M. K. Lim, "Using enhanced crow search algorithm optimization-extreme learning machine model to forecast short-term wind power," *Expert Systems with Applications*, vol. 184, Dec. 2021, doi: 10.1016/j.eswa.2021.115579.

[16] B. Czejdo, S. Bhattacharya, and C. Spooner, "Improvement of protein model scoring using grouping and interpreter for machine learning," *in 2019 IEEE 9th Annual Computing and Communication Workshop and Conference* (CCWC), Jan. 2019, pp. 0349–0353, doi: 10.1109/CCWC.2019.8666524.

[17] S. Z. Shogrkhodaei, S. V. Razavi-Termeh, and A. Fathnia, "Spatio-temporal modeling of PM2.5 risk mapping using three machine learning algorithms," *Environmental Pollution,* vol. 289, Nov. 2021, doi: 10.1016/j.envpol.2021.117859.

[18] Radovanović, S., Petrović, A., Delibašić, B., & Suknović, M. (2021, September 23). A fair classifier chain for multi-label bank marketing strategy classification. *International Transactions in Operational Research*, 30(3), 1320–1339. https://doi.org/10.1111/itor.13059

[19] Roy, N., Ahmed, R., Huq, M. R., & Shahriar, M. M. (2021). User-centric Activity Recognition and Prediction Model using Machine Learning Algorithms. *International Journal of Advanced Computer Science and Application12 (12)* https://doi.org/10.14569/ijacsa.2021.0121265.

[20] Duwairi, R. M., & Halloush, Z. A. (2022, June 1). Automatic recognition of Arabic alphabets sign language using deep learning. *International Journal of Electrical and Computer Engineering (IJECE),* 12(3),2996. https://doi.org/10.11591/ijece.v12i3.pp2996-3004.

[21] Hayder, I. M., Nabi Al Ali, G. A., & A. Younis, H. (2023, February 1). Predicting reaction based on customer's transaction using machine learning approaches. *International Journal of Electrical and Computer Engineering (IJECE),* 13(1), 1086. https://doi.org/10.11591/ijece.v13i1.pp1086-1096.

[22] B. Mytnyk, O. Tkachyk, N. Shakhovska, S. Fedushko, and Y. Syerov, "Application of artificial intelligence for fraudulent banking operations recognition," *Big Data and Cognitive Computing*, vol. 7, no. 2, May 2023, doi: 10.3390/bdcc7020093.

[23] Moro, S., Cortez, P., & Rita, P. (2014, June). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22–31. https://doi.org/10.1016/j.dss.2014.03.001
In-Text Citation: (Moro et al., 2014)

[24] J. Parmar, A. C. Patel, and M. Savsani, "Credit card fraud detection framework-a machine learning perspective*," International Journal of Scientific Research in Science and Technology*, pp. 431–435, Dec. 2020, doi: 10.32628/IJSRST207671.

[25] P. Sharma, S. Banerjee, D. Tiwari, and J. C. Patni, "Machine learning model for credit card fraud detection-a comparative analysis," *The International Arab Journal of Information Technology*, 2021, doi: 10.34028/iajit/18/6/6.

[26] M. Sudhakar and K. P. Kaliyamurthie, "A novel machine learning algorithms used to detect credit card fraud transactions," *International Journal on Recent and Innovation Trends in*

Computing and Communication, vol. 11, no. 2, pp. 163–168, Mar. 2023, doi: 10.17762/ijritcc.v11i2.6141.

[27] E. Ileberi, Y. Sun, and Z. Wang, "Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost," *IEEE Access*, vol. 9, pp. 165286–165294, 2021, doi: 10.1109/ACCESS.2021.3134330.

[28] P. S. Kumar, Preethika, Sivagami, Sridevipriya, and Vishali, "Credit card fraudulent detection using machine learning algorithm," *International Journal for Research in Engineering Application and Management*, pp. 445–449, Apr. 2020, doi: 10.35291/2454-9150.2020.0330.

[29] D. Kawade, S. Lalge, and D. M. Bharati, "Fraud detection in credit card data using unsupervised machine learning algorithm," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 5, pp. 5249–5256, May 2022,doi: 10.22214/ijraset.2022.42974.

[30] E. H. Muktafin, P. Pramono, and K. Kusrini, "Sentiments analysis of customer satisfaction in public services using K-nearest neighbors algorithm and natural language processing approach," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 1, pp. 146–154,Feb.2021,doi:10.12928/telkomnika.v19i1.1747.

[31] S. Redkar, S. Mondal, A. Joseph, and K. S. Hareesha, "A machine learning approach for drug-target interaction prediction using wrapper feature Selection and class balancing," *Molecular Informatics*, vol. 39, no. 5, May 2020, doi: 10.1002/minf.201900062.

[32] C. V. Sandeep and T. Devi, "A novel approach for bank loan approval by verifying background information of customers through credit score and analyze the prediction accuracy using random forest over linear regression algorithm," *Journal of Pharmaceutical Negative Results*, vol. 13, pp. 1748–1755,Jan. 2022, doi: 10.47750/pnr.2022.13.S04.211.

[33] H. Chen, "Prediction and analysis of financial default loan behavior based on machine learning model," Computational Intelligence and Neuroscience, pp. 1–10, Sep. 2022, doi: 10.1155/2022/7907210.

[34] F. Doko, S. Kalajdziski, and I. Mishkovski, "Credit risk model based on central bank credit registry data," *Journal of Risk and Financial Management*, vol. 14, no. 3, Mar. 2021, doi: 10.3390/jrfm14030138.

[35] S. Kokate and M. S. R. Chetty, "Credit risk assessment of loan defaulters in commercial banks using voting classifier ensemble learner machine learning model," *International Journal of Safety and Security Engineering*, vol. 11, no. 5, pp. 565–572, Oct. 2021, doi: 10.18280/ijsse.110508.

[36] F. Ahmed, "Ethical aspects of artificial intelligence in banking," Journal of Research in Economics and Finance Management, vol. 1, no. 2, pp. 55–63, Dec. 2022, doi: 10.56596/jrefm.v1i2.7.

[37] Elrefai, A. T., Elgazzar, M. H., & Khodeir, A. N. (2021, January 27). "Using Artificial Intelligence In Enhancing Banking Services". *2021 IEEE 11th Annual Computing and Communication Workshop andConference*(CCWC).https://doi.org/10.1109/ccwc51732.2021.9375993.

[38] Shady Abdelhadi, Khaled Elbahnasy, Mohamed Abdelsalam, "A Proposed Model To Predict Auto Insurance Clams Using Machine Learning Techniques", *journals of Theoretical and Applied Information* ,30th November 2020. Vol.98. No 22.