Bank Marketing Data Classification Using Optimized Voting Ensemble, Sine Cosine, and Genetic Algorithms

Marwa M. Eid[1], El-Sayed M. El-Kenawy[2], Abdelhameed Ibrahim[3], Abdelaziz A. Abdelhamid[4,5], Mohamed Saber[6]

[1] Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura 35712, Egypt
[2] Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology, Mansoura, 35111, Egypt
[3] Computer Engineering and Control Systems Department, Faculty of Engineering, Mansoura University, 35516, Mansoura Egypt
[4] Department of Computer Science, College of Computing and Information Technology, Shaqra University, Shaqra 11961, Saudi Arabia
[5] Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt
[6] Electronics and Communications Engineering Dep., Faculty of Engineering, Delta University for Science and Technology, Gamasa City, Mansoura, Egypt
Emails: marwa.3eeed@gmail.com; skenawy@ieee.org;afai79@mans.edu.eg;abdelaziz@su.edu.sa;abdelaziz@cis.asu.edu.eg; mohamed.saber@deltauniv.edu.eg

Abstract

Nowadays, the banking industry is no exception to the general trend of massive data production in all spheres of modern life. In this research, we analyze the categorization of marketing data from banks using a variety of machine learning techniques. The term "banking" refers to the supply of services by a bank to an individual consumer. The data was first compiled from the UCI Machine Learning repository and the Kaggle website. Phone-based banking marketing statistics are the focus of this data set. Python is utilized as the language of implementation, and the Machine Learning concept is employed for statistical learning and data analysis in this work. An improved prediction is the primary goal of machine learning's model-building phase. In order to classify the results, a supervised Naive Bayes algorithm is used to the data. The primary goal of the modeling effort is to characterize whether or not the consumer has chosen a term deposit. The bank should devote substantial time to returning phone calls from prospective customers. Accuracy, precision, recall, and F1 score were all evaluated as a consequence of this study in the direction of term deposit forecasting.

Keywords: Customer; bank marketing; machine learning; machine learning; metaheuristic optimization algorithms

1. Introduction

Banks are financial institutions that accept deposits from their customers and provide loans to them in exchange for interest. In order to better banking tactics and maintain solid client relationships, banks retain vast amounts of information about their consumers. A bank's customers are its most valuable asset. Direct marketing refers to the practice of advertising a product or service by making direct contact with its target audience, be it by traditional mail, electronic mail, human contact, mobile phone, or any other means of communication. Attracting new banking clients is a primary marketing goal[1]. The classification objective is to foretell if a consumer would subscribe the term deposit using

DOI: https://doi.org/10.54216/AJBOR.080202
Received: May 11, 2022 Accepted: December 09, 2022

data obtained from the UCI machine learning repository[2]. Machine learning technique for data analysis method and automates analytical building model to predict the accuracy of the bank customer data is used in this work. Python is used as the programming language[12], and its high-level, interpreter and extensive standard library are freely available source for all major platform from the Python web site. With the use of a classification technique, such as the Nave bayes classifier algorithm, we were able to get the most accurate results possible by measuring how well each instance in a dataset was defined by its properties. The bank should go after prospective clients who have spent a lot of time responding to bank calls. The overarching goal of this research is to construct a machine learning model employing a classification algorithm to forecast the correctness of the data, and to analyze and make predictions using an existing dataset in banking marketing to aid in successful decision making.

2. Literature Review

In [3], the author describes how she applied machine learning strategies to marketing efforts in the banking sector in order to analyse and anticipate trends using historical data. When it comes to banking marketing, the success rate is dependent on the outcome and the choice. Of this study, the authors employ a decision tree algorithm to forecast whether or not a consumer would sign up for a term deposit. This is one step in a multi-stage process that involves transforming raw data into effective decision-making knowledge and constructing a predictive model. According to Elsalamony et al., "all bank marketing campaigns are dependent on consumer huge data," yet it's hard for a human analyst to get useful insights from such a massive data pool. In this study, we applied the four most prominent data mining techniques—Multilayer Perception Neural Network (MLPNN), Nave Bayes (NB), logistic regression (LR), and decision tree—to improve campaign performance and pinpoint the factors responsible for its success[4]. In [5], the authors propose a data-driven method for forecasting the efficacy of bank telemarketing calls in pursuit of term deposits by employing a data mining technique. Considerable analysis of a wide range of factors pertaining to bank clients, economic and social characteristics, and products was included. Semi-automatic feature selection, implementation, and previous set reduction occurred during modelling.

Two metrics were used to evaluate the performance of four different types of data mining models—the super vector machine, a decision tree, a logistic regression model, and a neural network—with the latter proving to be the most successful. The decision tree, a knowledge extraction method, was then applied to the neural network to predict a number of important characteristics. We chose this model because we believe she will be an asset to our telemarketing efforts. In order to research consumer attributes and behaviour through efficient multi-channel communication, direct marketing is an interactive process[6] used by banks to cultivate positive relationships with their clientele. To boost client reaction to direct marketing campaigns is a primary objective of bank marketing, even more so than increasing profits, which may enhance customer satisfaction. Companies have been employing data mining techniques to reconstruct consumer profiles, as described in [7], and the banking industry has begun adopting a categorization approach to telemarketing. It was shown that using a combination of decision trees, random forests, and Naive Bayes to classify telemarketing leads was the most effective way to improve accuracy, precision, and recall. RapidMiner was utilized for both the experimentation and assessment processes, with pre-processing and normalization occurring prior to classifier evaluation. In the end, the results indicate that the decision tree is the most effective classifier for foreseeing client profile and behaviour.

3. Proposed Methodology

There are a number of steps involved in designing a system, from gathering data through processing it, training it, testing it, putting it into action, and predicting the outcome. Raw data is often unreliable because it lacks context or is contaminated. Before the data can be used for training a model, it must go through a process of pre-processing in which any errors or outliers are removed. Model performance may be improved with the help of feature engineering, which in machine learning involves activities such as feature selection and extraction. As an initial step in making a prediction, we typically train a model on a dataset so that it may use what it has learned to create predictions based on past experience. Overfitting is less likely to have occurred if a model fits both the training and test data, as the latter does not depend on the former. When it comes to machine learning, random forest is the algorithm of choice since it consistently produces high-quality outcomes. The process of preparing a model for use by feeding it training data and running it through an ML algorithm. At the

DOI: https://doi.org/10.54216/AJBOR.080202
Received: May 11, 2022 Accepted: December 09, 2022

end of the training phase, the model's accuracy is verified by comparison to the validation data (the test data).

Python is a powerful language that may be used for a wide variety of tasks since it is interpreted at a high level[12]. Python is a popular choice among developers because it can be picked up quickly, its syntax is straightforward, and it has a large number of available code libraries, making it simple to construct machine learning models. There are less lines of code in this program than in the competing language. There were a number of businesses that made use of In the fields of machine learning and data science, Anaconda is the most extensively used distribution of Python. Machine learning, a branch of AI[8], is a technique for analysing data and automating the creation of models; it "learns" from its past experiences, depending on the concepts it has been exposed to, and then uses that knowledge to make decisions with minimum human input. Accuracy and regularity are two of the primary concerns of machine learning. The majority of sectors, including banking, government, healthcare, retail, and transportation, are actively employing machine learning methods to evaluate massive data sets. Supervised learning is widely used in the field of machine learning today [9]. Data collection and data generation in supervised learning can be informed by prior knowledge. Algorithm learning from training data refers to the process by which a computer program figures out how to do a job by analysing data that has already been collected, in this case a mapping function between an input and an output variable. Classification and regression are the two main sub-fields of supervised learning. The unsupervised learning algorithm is used to train a machine with data that has not been labelled or classified, and the algorithm is given free reign to analyse the data however it sees fit. This algorithm focuses on uncovering patterns in data that have not been explicitly labelled, but it does not always produce accurate results[10]. Algorithms for unsupervised learning, which include things like grouping and association problems [11], aren't as precise as their supervised counterparts. In the fields of classification and regression, Random Forest [13] is a supervised, adaptable, and easy-to-use learning technique. This random forest aggregates the results of a large number of decision trees, uses these findings to make a final forecast, and uses an algorithm that reduces the danger of overfitting by choosing the prediction with the most votes. Reduces overfitting, improves accuracy, and may predict missing data are just a few of the benefits of the Random Forest method. Using the concept of independence across predictors, it is possible to determine the likelihood of an occurrence based on prior knowledge, making it the most successful categorization technique[14] available. Assuming that the existence of one feature in a class is independent to the presence of any other feature, the Naive Bayes classifier is naive.

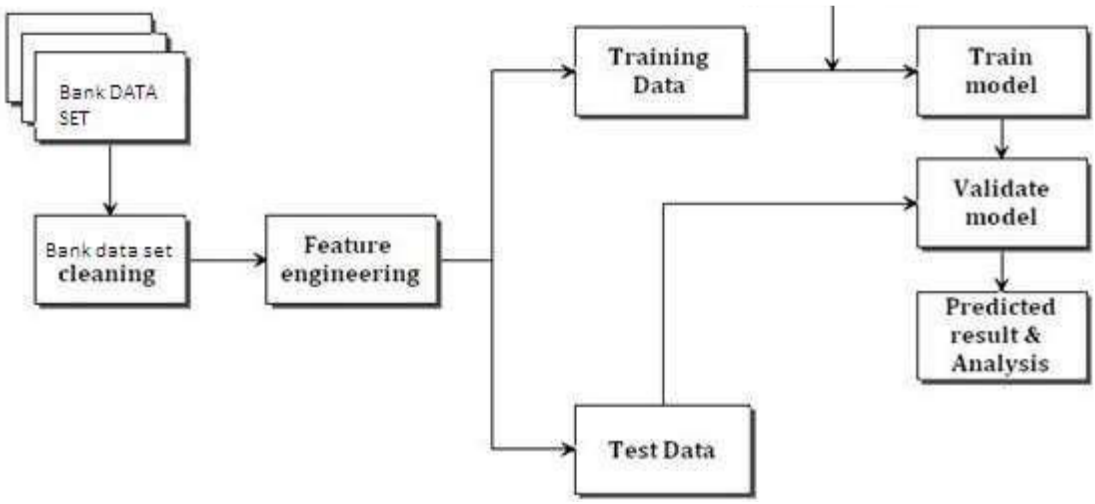

Figure 1: System Design.

DOI: https://doi.org/10.54216/AJBOR.080202
Received: May 11, 2022 Accepted: December 09, 2022

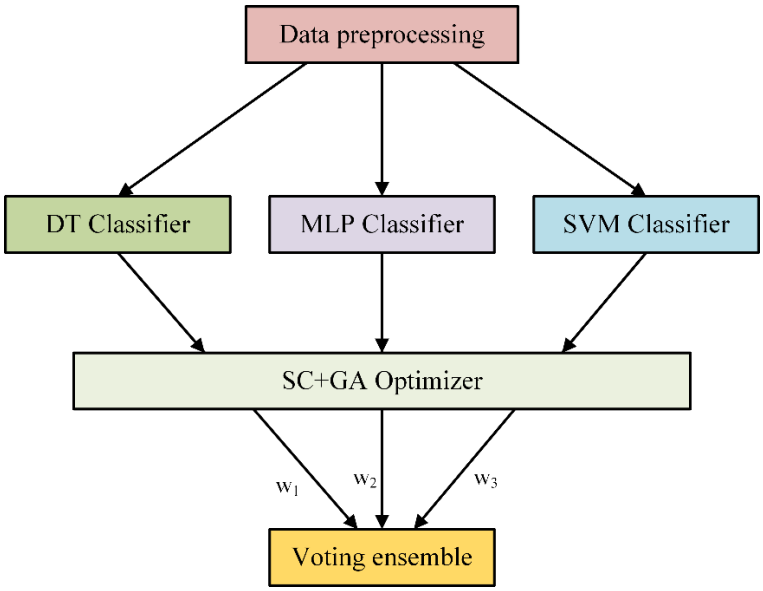

Figure 2: The proposed optimized voting ensemble approach

A. The Proposed Ensemble Model

The data we utilized to boat-train our classifiers after we let go is presented first, followed by a detailed explanation of the procedures we offer, and finally, the final product is displayed. This image was uploaded using three different methods of classification: Classifiers come in many forms, from decision trees (DT) to multilayer perceptron (MLP) to support vector machines (SVM) (SVM). As can be seen in Figure 2, we use the genetic algorithm (GA) method in combination with the sine cosine algorithm (SCA) to increase the importance of these classifiers' votes inside an ensemble model.

B. Support Vector Machines (SVM)

A relatively new advancement in the field of supervised machine learning is the support vector machine (SVM). If your dataset is small and has few outliers, this method will perform well. The idea behind hyper segmentation lanes is to organize data in meaningful ways. Along this hyperplane, the region has been partitioned into sectors, with each containing a unique set of data (Fig. 3). The two sets of data may be distinguished using a combination of hyperplanes. Ideally, we'd like to find the plane with the biggest safety margin. The two data points that are closest to the hyperplane that separates the two classes can be used to derive the margin. Support vector machines (SVMs) enhance the procedure by dividing the data into two equal halves and searching for the super planar with the biggest margin value. What we call "support vectors" are the data points that are closest to the hyperplane (Fig. 3). Identical to a hyperplane in that it is a linear surface that bisects space in two. A hyperplane, a one-dimensional subspace, is what you need to split your space in half.

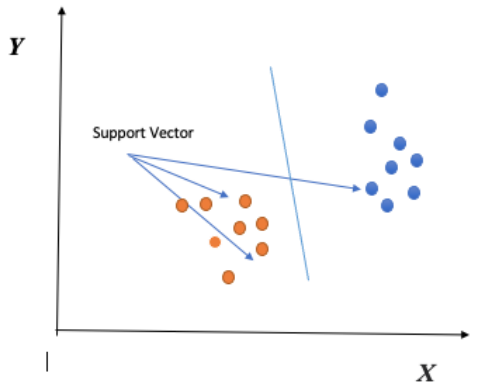

Figure 3: Structure of support vector machines.

DOI: https://doi.org/10.54216/AJBOR.080202
Received: May 11, 2022 Accepted: December 09, 2022

C. Multilayer Perceptron (MLP)

A multilayer perceptron neural network has been suggested. When separating the factors does not work out easily. A multilayer perceptron is built to solve this problem by including extra layers in a standard perceptron. The MLP network is a kind of feed-forward neural network, as shown in Figure 4, and it can include anywhere from one to several hidden layers. The network, which also contains n hidden neurons, receives n neutrons as input and returns n neutrons as output.

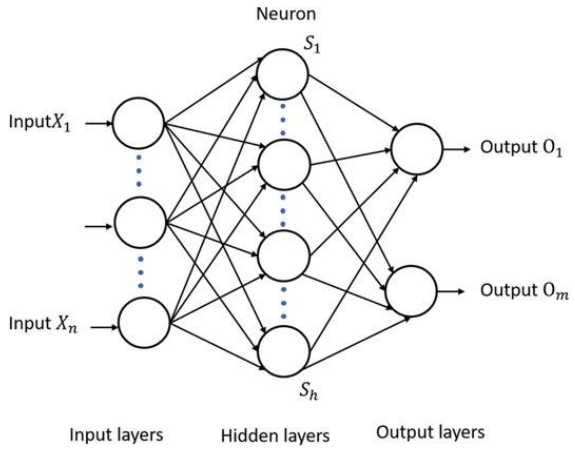

Figure 4: Structure of a multilayer neural network.

Our input matrix will have the form (batch size, number of attributes), and our weight matrices will have a single weight for each link to and from the hidden layer and the output layer as we will only be utilizing a single hidden layer. A matrix of (number of features, number of hidden neurons) represents the input layer, a matrix of (number of neurons, number of classes) represents the hidden layer, and a matrix of (number of features, number of hidden neurons, number of classes) represents the output layer (batch size, number of classes).

D. Decision Trees (DT)

In this diagram, each node in the tree stands for a property, and each branch shows a range of possible values for that attribute. Predictive value may be determined from a decision tree by following the paths of its nodes according to the values of the properties they represent. To correctly predict the end outcome, a tree method must be used as a starting point. A closer look is required at our data, its features, and the dummy values it contains. The following equations can help us find the best features for our tree by accounting for entropies and discriminative abilities. The decision tree from Figure 5 has been simplified and is shown below.

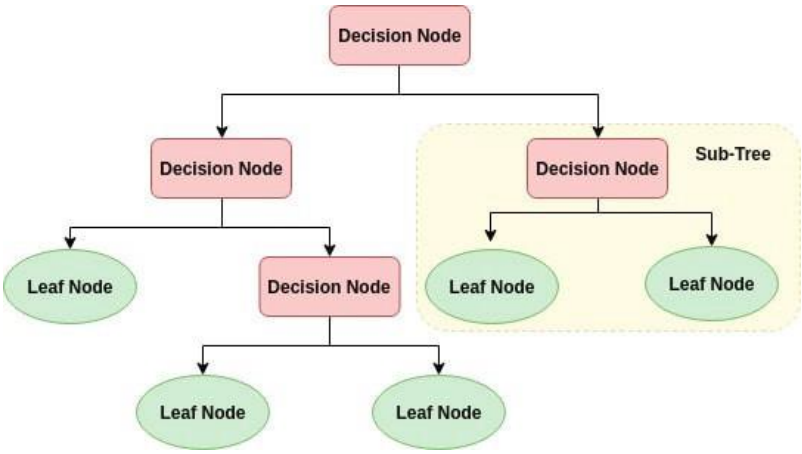

Figure 5: Structure of a decision tree.

In our data, the median age, for instance, serves as a clean boundary between two categories. Extra features were also tested using this technique. After that, we've identified the best places to divide, settling once and for all whether particular nodes in our tree would be on the left or the right. We

DOI: https://doi.org/10.54216/AJBOR.080202
Received: May 11, 2022 Accepted: December 09, 2022

utilize a certain column and value to choose where to make our data splits when building our tree of nodes (left child and right child of a node in a tree). In the recursive part of the algorithm, we employ the same methods as before to solve problems at progressively higher levels of the tree. The current node does not need to be transformed into a leaf node unless an inquiry is required.

E. Metaheuristic Optimization

In the domains of computer science and mathematical optimization, metaheuristics are used to locate, create, or choose a heuristic (partial search algorithm) that, despite missing data or insufficient computing power, may provide a feasible solution to an optimization issue. It would be challenging to collect a representative sample of the domain of possible solutions without the help of metaheuristics. Metaheuristics are often more adaptable than other optimization methods since they may be used in a variety of settings with little in the way of preparation or in-depth knowledge of the optimization issue at hand. There is no guarantee that the best solution to a problem class will be discovered through the use of metaheuristics, despite the fact that metaheuristics are being employed more frequently than traditional optimization approaches and iterative procedures. There are several well-known metaheuristics, stochastic optimization being only one of them, whose final solution may vary depending on the values of the random variables. In combinatorial optimization, metaheuristics can be more efficient than optimization algorithms, iterative techniques, or even simple heuristics when used to generate hypotheses regarding the ideal solution. Thus, they may offer novel strategies for addressing optimization issues. There are a lot of articles out there on the topic. Most metaheuristics publications have an experimental tone since they report the author's own experiences with implementing the algorithm. However, in addition to empirical data, there are formal theoretical results that provide light on concerns of convergence and the viability of reaching a global optimum. Recent research has offered a number of alternative metaheuristic strategies, all of which have the potential to significantly impact the area as a whole. Previous studies on this topic suffered from a lack of precision in their wording, an inability to thoroughly cover crucial topics, a lack of proper methodology, and a lack of appropriate citations.

4. Results

Predictions of term deposits are made using the Customer bank dataset. The UCI machine learning repository makes this dataset available to the general audience. There's data on customers there. The predictions in the dataset can be either Yes or No. Each of the 16 input features contributes to the single output. Using metrics like accuracy, precision, recall, and f1-score, we find that after applying the Supervised Nave Bayes algorithm for classification purposes, the approach provides 82.65 accuracy for the dataset.

Table 1: Attributes of bank dataset

ID	Attributes	Type	Values	Descriptions
1	Age	Numeric	Real	Age at the contact date (≥18)
2	Job	Categorical	Admin, Unknown, Unemployed, Management, Housemaid, Entrepreneur, Student, Blue-collar, Self-employed, Retired, Technician, Services	
3	Marital	Categorical	Married, Divorced, Single, widowed	
4	Education	Categorical	Unknown, Secondary, Primary, Tertiary	
5	Default	Binary	Yes, No	Yes or No
6	Balance	Numeric	Real	In euro currency
7	Housing	Binary	Yes, No	Yes or No
8	Loan	Binary	Yes, No	Yes or No
9	Contact	Categorical	Unknown, Telephone, Cellular	
10	Day	Numeric	Real	Referring to when the contact was made
11	Month	Categorical	Jan, Feb, mar,..., Nov, Dec	
12	Duration	Numeric	Real	Of the contact (in seconds)
13	Campaign	Numeric	Real	
14	Pday	Numeric	Real	
15	Previous	Numeric	Real	
16	Poutcome	Categorical	Unknown, Failure, Success	

DOI: https://doi.org/10.54216/AJBOR.080202
Received: May 11, 2022 Accepted: December 09, 2022

Table 2: Classification results using the proposed method compared to other methods

	Accuracy	Sensitivity	Specificity	Pvalue	Nvalue	F-score
NN	0.9010	0.1250	0.9677	0.2500	0.9278	0.1667
SVM	0.8617	0.0909	0.9639	0.2500	0.8889	0.1333
DT	0.7969	0.0909	0.9434	0.2500	0.8333	0.1333
SC+GA	0.9381	0.2500	0.9677	0.2500	0.9677	0.2500

Table 3 displays the statistical evidence for the superiority of the voting ensemble classifier. The given optimal voting ensemble yields significantly enhanced performance.

Table 3: Statistical analysis of the results recorded by the proposed method

	NN	SVM	DT	SC+GA
Number of values	10	10	10	10
Minimum	0.901	0.8617	0.7869	0.9381
25% Percentile	0.901	0.8617	0.7969	0.9381
Median	0.901	0.8617	0.7969	0.9381
75% Percentile	0.901	0.8642	0.7969	0.9381
Maximum	0.911	0.8817	0.8197	0.9381
Range	0.01	0.02	0.03281	0
10% Percentile	0.901	0.8617	0.7879	0.9381
90% Percentile	0.91	0.8807	0.8174	0.9381
95% CI of median				
Actual confidence level	97.85%	97.85%	97.85%	97.85%
Lower confidence limit	0.901	0.8617	0.7969	0.9381
Upper confidence limit	0.901	0.8717	0.7969	0.9381
Mean	0.902	0.8647	0.7982	0.9381
Std. Deviation	0.003162	0.006749	0.008192	0
Std. Error of Mean	0.001	0.002134	0.002591	0
Lower 95% CI of mean	0.8997	0.8599	0.7923	0.9381
Upper 95% CI of mean	0.9043	0.8695	0.804	0.9381
Coefficient of variation	0.3506%	0.7806%	1.026%	0.000%
Geometric mean	0.902	0.8647	0.7981	0.9381
Geometric SD factor	1.003	1.008	1.01	1
Lower 95% CI of geo. mean	0.8997	0.8599	0.7923	0.9381
Upper 95% CI of geo. mean	0.9042	0.8695	0.8039	0.9381
Harmonic mean	0.902	0.8647	0.7981	0.9381
Lower 95% CI of harm. mean	0.8997	0.8599	0.7924	0.9381
Upper 95% CI of harm. mean	0.9042	0.8694	0.8039	0.9381
Quadratic mean	0.902	0.8647	0.7982	0.9381
Lower 95% CI of quad. mean	0.8997	0.8598	0.7923	0.9381
Upper 95% CI of quad. mean	0.9043	0.8696	0.8041	0.9381
Skewness	3.162	2.277	2.155	
Kurtosis	10	4.765	6.862	
Sum	9.02	8.647	7.982	9.381

DOI: https://doi.org/10.54216/AJBOR.080202
Received: May 11, 2022 Accepted: December 09, 2022

The Wilcoxon signed-rank test is used to evaluate the proposed method against the alternatives. The data collected throughout the course of this investigation is shown in Table 4. The p-values presented in the table serve as evidence.

Table 4: Wilcoxon signed rank test of the recorded results of the proposed method

	NN	SVM	DT	SC+GA
Theoretical median	0	0	0	0
Actual median	0.901	0.8617	0.7969	0.9381
Number of values	10	10	10	10
Wilcoxon Signed Rank Test				
Sum of signed ranks (W)	55	55	55	55
Sum of positive ranks	55	55	55	55
Sum of negative ranks	0	0	0	0
P value (two tailed)	0.002	0.002	0.002	0.002
Exact or estimate?	Exact	Exact	Exact	Exact
P value summary	**	**	**	**
Significant (alpha=0.05)?	Yes	Yes	Yes	Yes
How big is the discrepancy?				
Discrepancy	0.901	0.8617	0.7969	0.9381

Comparison of the optimum voting ensemble classifier's results to those of the baseline models is depicted graphically in Figure 6. For an example of the enhanced efficacy of the proposed method is shown superior when compared to the other methods.

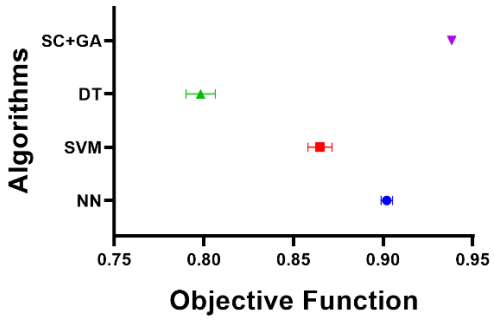

Figure 6: The accuracy of the proposed method compared to other methods.

5. Conclusion

There is a constant stream of data creation in the banking industry, and that data may be mined for useful insights. Predicting whether or not a consumer will sign up for a term deposit is the primary focus of this effort. This paper's research makes use of a bank dataset for classification purposes, which was obtained either from the UCI machine learning repository or the Kaggle website. The end product was good once all the work was done.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] https://www.technofunc.com/index.php/domain-knowledge/banking-domain/item/what-is-a-bank

[2] archive.ics.uci.edu/ml/datasets/Bank+Marketing.

[3] Mihova, Yana, Knowledge creation in banking marketing using machine learning techniques, 2019.
[4] El-sayed M. El-kenawy, Marwa M. Eid, Abdelhameed Ibrahim, Anemia Estimation for COVID-19 Patients Using A Machine Learning Model. Journal of Comnputer Science and Information Systems, 2(1) ,1-7, 2021.
[5] Moro, S et al.. A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31, 2014.
[6] Miguéis, V.L., Camanho, A.S. & Borges, J., Predicting direct marketing response in banking: comparison of class imbalance methods. Service Business, 11, 831–849, 2017.
[7] S. Palaniappan, A. Mustapha, C. F. M. Foozy, and R. Atan, Customer profiling using classification approach for bank telemarketing. JOIV: International Journal on Informatics Visualization, 1(4), 214–217, 2017.
[8] https://www.sas.com/en_in/insights/analytics/machine-learning.html
[9] https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
[10] https://www.expert.ai/blog/machine-learning-definition
[11] https://www.geeksforgeeks.org/supervised-unsupervised-learning/
[12] Akshansh Sharma et all. (2020). Python: The Programming Language of Future, IJIRT, 6 Issue 12.
[13] W.T. Aung, K.H. Hla, Random forest classifier for multi-category classification of web pages, in IEEE Asia-Pacific Conference on Service Computing, Biopolis, Singapore, 372–376, 2009.
[14] Kaviani, Pouria & Dhotre, Sunita, Short Survey on Naive Bayes Algorithm. International Journal of Advance Research in Computer Science and Management. 04(11), 2017.

DOI: https://doi.org/10.54216/AJBOR.080202
Received: May 11, 2022 Accepted: December 09, 2022