Received: 4 June 2023 | *Revised:* 22 August 2023 | *Accepted:* 25 August 2023
DOI: 10.37917/ijeee.20.1.23 *Vol. 20 | Issue 1 | June 2024*

Iraqi Journal for Electrical and Electronic Engineering
Original Article

Open Access
IJEEE University of Basrah College of Engineering

Digital Marketing Data Classification by Using Machine Learning Algorithms

Noor Saud Abd*[1], Oqbah Salim Atiyah[1], Mohammed Taher Ahmed[1], Ali Bakhit[2]
[1]Department of Computer science, College of Computer Sciences, University of Tikrit, SalahAldeen, 34001, Iraq.
[2]Higher Institute of Computer Science and Information Systems, Culture & Science City, Giza, 12511, Egypt.

Correspondance
*Noor Saud Abd
Department of Computer science,
College of Computer Sciences,
University of Tikrit, SalahAldeen, 34001, Iraq.
Email: noor.s.abd@tu.edu.iq

Abstract
Early in the 20th century, as a result of technological advancements, the importance of digital marketing significantly increased as the necessity for digital customer experience, promotion, and distribution emerged. Since the year 1988, in the case when the term "Digital Marketing" first appeared, the business sector has undergone drastic growth, moving from small startups to massive corporations on a global scale. The marketer must navigate a chaotic environment caused by the vast volume of generated data. Decision-makers must contend with the fact that user data is dynamic and changes every day. Smart applications must be used within enterprises to better evaluate, classify, enhance, and target audiences. Customers who are tech-savvy are pushing businesses to make bigger financial investments and use cutting-edge technologies. It was only natural that marketing and trade could be one of the areas to move to such development, which helps to move to the speed of spread, advertisements, along with other things to facilitate things for reaching and winning customers. In this study, we utilized machine learning (ML) algorithms (Decision tree (DT), K-Nearest Neighbor (KNN), CatBoost, and Random Forest (RF) (for classifying data in customers to move to development. Improve the ability to forecast customer behavior so one can gain more business from them more quickly and easily. With the use of the aforementioned dataset, the suggested system was put to the test. The results show that the system can accurately predict if a customer will buy something or not; the random forest (RF) had an accuracy of 0.97, DT had an accuracy of 0. 95, KNN had an accuracy of 0. 91, while the CatBoost algorithm had the execution time 15.04 of seconds, and gave the best result of highest f1_score and accuracy (0.91, 0. 98) respectively. Finally, the study's future goals involve being created a web page, thereby helping many banking institutions with speed and forecast accuracy. Using more techniques of feature selection in conjunction with the marketing dataset to improve diagnosis.

Keywords
CatBoost, Decision Tree, Digital Marketing, K-Nearest Neighbor (KNN), Machine learning, Random Forest (RF).

I. Introduction

Early in 20th century, as a result of technology advancements, the importance of digital marketing significantly increased as the requirement for digital customer experience, promotion, and distribution emerged [1]. Since the year 1988, in the case when the term "Digital Marketing" first appeared, the business sector has undergone a drastic growth, moving from small startups to massive corporations on a global scale.

Digitalization, information and communication technology, ML, AI, and robots are causing dramatic change in the world we live in. This essay explores the relation between digital marketing and AI, both current and future, while simultaneously suggesting methods for AI involvement in app development [2]. Digital marketing, a legitimate subfield

of marketing science, was able to benefit firms and boost customer involvement through electronic offerings. The emergence of the digital technology made it easier for the industries to monitor their production, channel distributions, branding, promotion, and advertisement processes. Business managers might make more informed decisions depending upon the obtained data, interactive customer experiences, and digital overview of data and sales. Customers' experiences and journeys are becoming incredibly difficult as a result of excessive amount of data that is produced every day [3]. Organizations spend a lot of information to compensate for a lack of data or unmapped potential customers. The overwhelming amount of data produced has created a confusing landscape that marketers should navigate. Decision-makers should handle the reality that users' data change on day-to-day basis [4].

In order to effectively classify, analyze, optimize, and target audiences, enterprises must employ smart applications. Customers that are tech savvy encourage industries to make larger financial expenditures and use more complex solutions. Marketers should determine their goals and look for cutting-edge technology solutions in light of a very complicated data landscape [5]. The business world has the ability of installing the smart apps which have direct impact upon decision-makers and marketing sector. Customer action predictions based upon the dependent variables of interest may result from intelligent data-driven models. AI, ML, DL, and DM can support marketing science. The advantages of AI include data classification, user profiling, optimized targeted audiences, content optimization, predictive models, and optimizations of search engine ranking variables [6].

Digital marketing underwent a period of transition because of the scientific developments in AI, computational analysis, and DM, which resulted in shifting from a completely data-driven method to a new method combining data and knowledge-based systems of decision-making [7]. Making decisions can appear like a difficult effort when trying to create a qualitative marketing decision because of factors including knowledge gained from professional experience, judgment calls, specialized knowledge, and a multidimensional environment that is demanding and always changing. The volume of data produced every day provides marketers with the chance to build, construct, and use apps that combine scientific multifactor data with a goal of developing smart knowledge-based decision-making patterns [8]. We will briefly discuss some prior work on the topic before moving on to the algorithms and approaches employed in this research, followed by the findings and conclusions.

II. Related Work

This section summarizes some prior research and describes the various methods utilized to categorize customer purchases in digital marketing:

- Gao, W., & Ding, Z. (2022) [9].

  In comparison to other linear classifiers, it has a number of benefits, including higher prediction accuracy and reduced generalization error. The conventional RF and a few data processing algorithms are examined and researched in this work, and the RF is used to investigate and research difficulties related to class imbalance and feature selection. By researching feature selection techniques, we may balance feature relevance and strength throughout feature selection, enhancing the influence of the final model categorization. The RF model technique was enhanced to address the unbalanced problem through research and experimentation on the RF unbalanced data classification challenge. Following calculations and empirical study, it was discovered that in the case when two samples have been taken along with the bare minimum of the samples that are needed for a node to divide by various numbers, the best results are acquired. The average F1 evaluation value is 0.1038, thus splitting a node is necessary; the impact of randomized forests is the best application of Gini index for various specifications, and mean value of its F1 evaluation is 0.1033; for analyzing impact of RF with various numbers of trees, 7 - 10 DTs are best, and F1 rating is the best, mean-e is 0.101750 on average.

- Gkikas, D. C., Theodoridis, P. K., et al, (2022), [10].

  Consumer behavior can be predicted using this approach on both a physical and digital level. Through the single encapsulation approach, also referred to as the GA encapsulation approach, it integrates DTs and genetic algorithms (GAs). Depending on the goals of the study, consumer survey data was gathered and categorized. Software developed by GA Embedding was shown to function remarkably well, with classification accuracy over 90%. Household size and monthly income categories allow for a more accurate identification of particular subgroups of genes which affect decision-making when it comes to gender. Those categories were linked to a particular set of characteristics, offering a clear road map for choosing marketing strategies.

- Li, Z. (2022)) [11].

  They put forward an improved algorithm of XGBoost based on Bayesian optimization parameters, which can improve the efficiency of digital marketing communication and enhance the social influence of digital marketing. The data that have been crawled through the network are processed, and the data set is modeled, in

which (60%) of the data are randomly selected as training sets, (40%) as test sets, and accuracy and recall are used as performance measurement methods of the model. through verification, the improved XGBoost algorithm based on Bayesian parameter selection is obviously superior to logistic regression in accuracy and recall, and its computational efficiency is also obviously superior to logistic regression, which shows certain advantages in data processing and analysis.

- De Mauro, A., Sestino, A., (2022) [12].

  The propose was a taxonomy of ML use cases in marketing based on a systematic review of academic and business literature. We have identifed 11 recurring use cases, organized in 4 homogeneous families which correspond to the fundamentals leverage areas of ML in marketing, namely: shopper fundamentals, consumption experience, decision making, and fnancial impact. We discuss the recurring patterns identifed in the taxonomy and provide a conceptual framework for its interpretation and extension, highlighting practical implications for marketers and researchers.

- Angelina, J. J. R. (2023, May) [13].

  This research study has used Cat Boost Classifier to predict the customer behavior, retain the customers and increase the company's benefit. The proposed model provides a prediction accuracy of 95%, when compared with the existing methods.Quality of Service (QoS) is now attracting an increasing number of clients towards a company. Meanwhile, the ability to provide clients with technologically better QoS is highly demanding in recent times. ON the other hand, effective customer interactions may assist the organization to gain new clients, retain existing ones, and improve customer retention by bringing in more income.

III. Digital Marketing

Via internet, digital marketing is able to alter consumer perceptions in addition to increasing the on-line sales. The digital era has offered consumers the chance for expressing themselves and voice their opinions, giving them a power of choice and impact at the same time. A user can search for the information, products, or services, and brands has the ability to communicate with the customers in real time. Businesses were able to broaden their customer bases thanks to digital marketing.

Businesses and consumers have benefited from digital marketing [14]. Through recommending new goods for digital marketing that the consumer has little experience with and that are relevant to their present needs, the recommendation system for digital marketing addresses the issue of information overload. The data regarding users, ongoing transactions, available items, and several other sorts of data stored in the current database environment are used by the digital marketing recommendation system to make recommendations [15]. Users might accept or reject the information that is offered by digital marketing, and they could also give implicit or explicit feedback over time, both of which are very helpful to the systems that make recommendations for digital marketing [16]. As a result, all user feedback data could be kept in the related database of digital marketing suggestions, allowing the system to develop new recommendations for digital marketing in the case when the user exhibits relevant behavior in the future. A major issue in digital marketing is the classification problem. In order to categorize the unknown data, it is necessary to find a collection of models which could describe the shared data regarding all known data. This is referred to as "classification." The classification model that reflects a classification knowledge body, is typically created when a classification problem needs to be solved. Researchers have put forth a large number of classification algorithms from many domains with various purposes in an effort to solve the classification problem [17].

Those classification algorithms could be used to solve a variety of classification problems in digital marketing. The application and viability of the algorithms of classification in the digital marketing will be discussed in length, along with the marketing scenarios from the real-world work, in this article, which will also introduce a number of frequently utilized classification algorithms thoroughly. The foundational framework for marketing activities is provided by the theory of digital marketing, which focuses on how to employ RFs to enhance data analysis capabilities and evaluate marketing data effectively. This study provides a detailed introduction to the theory and definition of classification problem, as well as the common problem of classification in the area of the digital marketing. It then employs an RF method so as to resolve the common problem of classification [18].

IV. Material And Method

In this section will show classification in machine learning and the algorithms used in this paper. With some details.

Classification Algorithms

There are numerous ML algorithms that have been modeled for classification. Here are some of them will be explained:

A. Decision Tree

A DT can be defined as a top-to-bottom classification of a tree structure. Its core concept is purity classification. Purity denotes the ability to totally isolate the target variables, i.e., y

could just equal 0 throughout classification, or 1. One of the oldest DM algorithms is the DT. It displays the classification and decision-making process in a straightforward, understandable, and highly comprehensible tree form [19]. Classification trees are utilized in this study. How should one evaluate the quality of leaf nodes that are produced through the splitting of the root node into leaf nodes in the decision-making process? Commonly, Gini index (which is referred to as the Gini impurity as well) is utilized to judge. Gini index specifies the likelihood that a sample which is randomly selected within the set will be incorrectly classified; consequently, the lower the Gini index, the lower the likelihood that the selected sample within the set will be incorrectly classified. To put it another way, the higher the set's purity, the more impure the set is, and the more -at is, the more impure the split leaf node would be the following is a mathematical formula [20]:

$$Gini(P) = \sum_{p=1}^{k} P_k(P_k - 1) + P_k^2 \tag{1}$$

P denotes the likelihood that a randomly chosen sample falls under K category among them. *Gini* index purity is determined after root node has been split to 2 leaf nodes. Those two leaf nodes could be indicated as leaves in the case when the purity and *Gini* values are both sufficiently low. No longer are the nodes categorized. The two leaf nodes that have been mentioned above will be utilized as a new set and split repeatedly in a case when the purity is high, doing so until the purity is low enough for meeting the criteria [21].

B. Random Forest

Explain the DT after it has been indicated. A collective learning technique called RF uses a large number of DTs to create regression, classification, and other tasks. In other words, the forest is made up of several DTs, none of which are connected to one another. In the case where a new sample enters the forest after the RF has been built, each DT in a forest performs decision classification in order to determine the class that is chosen the most to anticipate the class that this new sample belongs to. The phenomena of repeated sampling could happen because the RF will sample input dataset randomly with replacement of columns and rows [21]. Each DT needs m sample sets to be trained, assuming there are m DTs. Full samples will disobey local sample restrictions and make the model less generalizable, hence it isn't suggested to utilize full samples for the training of m trees. The prediction classification is then generated when m DTs are trained using the n samples that were extracted with replacement. The chain rule is typically used to calculate many random variables, as seen in Eq. (2).

$$P(x_1, x_2 \ldots x_n) = p(x_1)U(p_x(x_i|x_1 \ldots x_{i-1}) \tag{2}$$

Users do not require a great deal of mathematical or statistical expertise to use RF because of its high operational efficiency, easy to grasp nature, and straightforward implementation. Along with the benefits already described, RF also has a strong anti-overfitting capability and could be utilized simultaneously for the problems of regression and classification. Yet, compared to the classification problem, the RF application effect in regression problem is less favorable. When addressing the regression problem, it cannot provide predictions that extend beyond data range of a training sample set since it is unable to produce a continuous output. Data have a tendency to overfit. The impact will be lessened in the case forecasting for low-dimensional data sets because the benefit of RF is its ability to handle high-dimensional and unbalanced data sets [22].

C. K-Nearest Neighbor (KNN) Algorithm

To determine which k training samples in the training set are the nearest to target object, the K-NN is suggested. Assign the dominating category to the target object after identifying it from the k training examples, in which k represents the number of training samples. Since all samples in a feature space are categorized into the same class, which includes the samples that are k most closely related to each other, K-NN algorithm's main mechanism is that all of the samples have the same features while being categorized into that class. When making a classification decision, the approach solely considers the category of the sample or samples that are nearest in proximity when identifying the class to which a sample belongs. In category decision making, K-NN is also only applicable to a relatively small set of the subsequent samples [23]. K-NN must determine how far the predicted data point is from known data point in order to choose nearest k labeled data points, $y_1, y_2, \ldots, y_k$, in which y_1 denotes known data point that is closest to predicted point, y_2 denotes known data point which is second nearest to forecasted point, etc. As a result, Eq. (1) could be used to implement K-NN regression for the short-term load predictions.

$$S_i = \frac{1}{k} \sum_{j=1}^{k} s_{y_j} \tag{3}$$

s_i denotes i_{th} predicted value, which is average value of $s_{y_j}(j = 1, 2, \ldots, k)$; s_{y_j} denotes forecasted value of jth nearest known data point (y_j) [24].

It is a very simple supervised learning-based ML algorithm. K-NN places the new state in a category which is most similar to available categories by assuming similarity between the available states and the new state/data. Each piece of information is stored by K-NN, which uses similarity to categorize new data point. This means that utilizing the K-NN algorithm, new data could be quickly and accurately categorized into

a good group class [25]. Although the K-NN could also be utilized to solve regression problems, it is most frequently employed to solve classification problems [26]. Since K-NN is a non-parametric algorithm, no assumptions are made about the underlying data. The reason it does not learn from training dataset right away is since it saves dataset and performs an action on the data set when it comes time to classify. This algorithm is also referred to as lazy learner algorithm. With regard to the training phase, K-NN simply retains the training data set, and in the case where it receives new data, it performs the classification of that data into a category which is similar to training data [27].

D. CatBoost Algorithm

The CatBoost algorithm is a member of the family of GBDT machine learning ensemble techniques. Since its debut in late 2018, researchers have successfully used CatBoost for machine learning studies involving Big Data. The term CatBoost is an acronym that stands for ”Category” and “Boosting.”It is used for search, recommendation systems, personal assistants, self-driving cars, weather prediction, and many other tasks. One of the many unique features that the CatBoost algorithm offers is the integration to work with diverse data types to solve a wide range of data problems faced by numerous businesses. Not just that, but CatBoost also offers accuracy just like the other algorithm in the tree family [27]. According to the CatBoost documentation, CatBoost supports numerical, categorical, and text features but has a good handling technique for categorical data. The CatBoost algorithm has quite a number of parameters to tune the features in the processing stage.”Boosting” in CatBoost refers to the gradient boosting machine learning. Gradient boosting is a machine learning technique for regression and classification problems [28]. The CatBoost algorithm utilizes efficient modified target-based statistics to appropriately handle the categorical features during training time, thus saving considerable computational time and resources. Another important aspect of the CatBoost algorithm is its ordered boosting mechanism. In traditional GBTs, all the training samples are provided to construct a prediction model after executing several boosting steps. This approach causes a prediction shift in the constructed model, which consequently leads to a special kind of target leakage problem. The CatBoost algorithm avoids the stated issue by utilizing the ordered boosting framework. Furthermore, contrary to the conventional learning classifiers, the CatBoost algorithm eloquently handles the overfitting issue by using several permutations of the training dataset; hence it turns out to be the key motivation behind utilizing its intelligence in the current study [29].

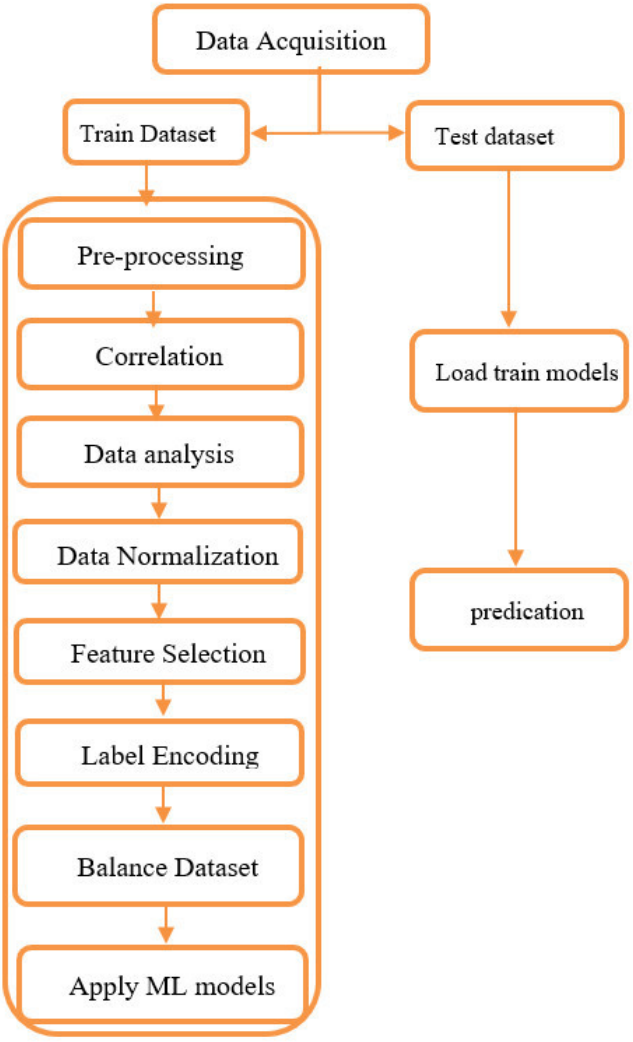


Fig. 1. System work flow.

V. Proposed System Framework

The mechanism by which this study is carried out is depicted in the figure below. All of the experiments were carried out using the Python programming language. The flow diagram above demonstrates how the experiment will be conducted with the use of python code on the dataset that was downloaded from the data world. The suggested system generally consists of a set of steps. According to Fig. 1, each level consists of a set of smaller steps: All tables should be numbered with numerals. Every table should have a caption. Headings should be placed above tables, left justified. Only horizontal lines should be used within a table, to distinguish the column headings from the body of the table, and immediately above and below the table. Tables must be embedded into the text and not supplied separately.

A. Dataset Description

Information relates to the direct marketing data regarding a Portuguese banking organization. On the phone conversations, marketing campaigns have been based. For determining whether the customer will subscribe to the product (bank term deposit), more than one contact with the same client was frequently necessary. The classification goal is to predict if the client will subscribe to a term deposit (variable y). folder Initially, the datasets consisted of two separate files containing each of them [30]:

- train.csv: 45211 rows and 17 columns that have been ordered by the date.
- test.csv: 4,521 rows and 17 columns with 10% of samples (45,211), chosen in a random manner from the train.csv.

B. Pre-processing

This step is in the training phase, and also includes other phases that we will explain below:

1) Descriptions of the Column

Bank client data with column descriptions after viewing them in python:

1. age (numerical)
2. marital: marital status (categorical: ”single”, ”married”, ”divorced”; note: ”divorced” entails widowed as well as divorced)
3. job: job type (categorical: ”admin.”, ”unemployed”, ”unknown”, ”housemaid”, ”student”, ”management”, ”entrepreneur”,”self-employed”, ”blue-collar”, ”retired”, ”services”or”technician”)
4. default: has credit in default? (binary: ”yes” or ”no”)
5. education (categorical: ”primary”, ”unknown”, ”tertiary”, or”secondary”)
6. housing: has housing loans? (binary: ”yes” or ”no”)
7. balance: average annual balance value, in euros (numerical)
8. contact: type of the contact communications (categorical: ”telephone”,”unknown”,or”cellular”)
9. loan: has personal loans? (binary: ”yes” or ”no”) # associated to last contact of current campaign:
10. month: last contact month of the year (categorical: ”jan”, ”feb”,”dec”)
11. day: last contact day in the month (numerical)
12. duration: last contact duration, in sec. (numerical) other.attributes:
13. p_days: number of the days which passed by after a client has been last contacted from a preceding campaign (numerical, -1 indicates the fact that the client has not been contacted earlier)
14. campaign: number of the contacts that are carried out throughout this campaign and for the client (numerical, including,the:last:contact)
15. p_outcome: result of previous marketing campaign (categorical: ”failure”, ”other”, ”success”, and”unknown”)
16. previous: number of the contacts that are carried out before this campaign and for that client (numerical)

    Output variable (wanted target):

17. y - has client subscribed a term deposit? (binary: ”yes” or ”no”). Missing Attribute Values: None. Banking Data-set of various customers to predict whether or not they will convert, was taken from Kaggle on November 15, 2022, With Note, updated about two years prior to this date [30].

2) Data Analysis

There are many things involved in data analysis, but we will explain here the most important two stages in it.

1. Check Missing Values: This phase is important, to check the data if has missing data to treatment it. After checking the data, it was not found missing data, as the Table I.
2. Check Data Types: After checking the missing values and processing them if it is found, must be known types of data to encode the categorical data into numeric because some algorithms cannot deal with it mathematically until after converting it. Table I appears there are numeric and categorical data.

C. Correlation

Correlation represents an indication concerning changes between a pair of variables. The correlation matrix is plotted to show the variable that has a low or high correlation compared with the other variable [31]. Fig. 2 the correlations heat map between features, where the scale measures the correlation degree among all features, and the correlation values are in the range [-1,1]. Score (1): This value refers to the correlation that is completely directly between the two features. Score (0):

TABLE I.
MISSING VALUE AND DATA TYPES

Features	Missing Values	Data Types
age	0	object
job	0	object
marital	0	object
education	0	object
default	0	int64
balance	0	object
housing	0	object
loan	0	object
contact	0	int64
day	0	object
month	0	int64
duration	0	int64
campaign	0	int64
pdays	0	int64
previous	0	object
poutcome	0	object
y	0	object

This value refers to the correlation between the two features that are absent. Score (-1): This value refers to the correlation that is inversely proportional between the two features.

D. Data Normalization

Normalization is a technique that is often applied as part of preparing data for machine learning, in preprocessing stage. Normalization has the objective of scaling down the features to some similar scale. Which improves model's functionality and training stability, and increases the data's integrity and accuracy [32]. We used MinMaxScaler function, which individually scales each feature where the values have a given maximum and minimum value, with a default value of 1 and 0.

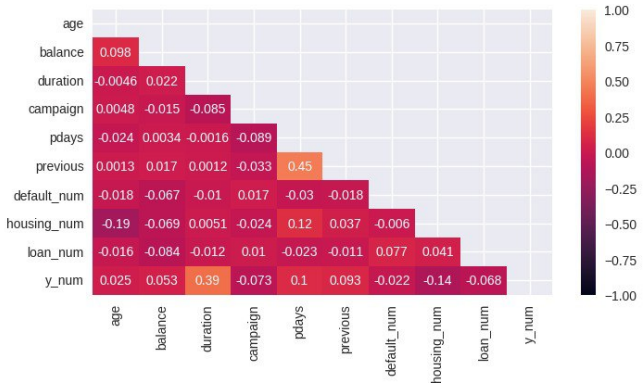


Fig. 2. The correlations heat map between features.

TABLE II.
THE FEATURE SELECTION OF NUMRECL DATA

Features	Without ANOVA	With ANOVA
age	0.684818	- 8.826
balance	8.360308	- 2.521
duration	3.144318	0.0
campaign	4.898650	- 1.012347
pdays	2.615715	- 3.790553
previous	41.846454	- 7.80183
default_num	7.245375	- 1.86635
housing_num	-0.224766	- 2.62192
loan_num	1.852617	- 9.7936575
y_num	2.383480	- 8.825643

E. Feature Selection

Feature selection is a step in the data pretreatment process, which is the longest step in any ML pipeline. You can approach it more methodically in a way that is conducive to the aid of these ML strategies. You will be better able to decipher the features. The procedure is improved by using feature selection, which also speeds up the ML algorithm's learning process. By choosing the most crucial variables and removing irrelevant or redundant variables, the algorithms' predictive power is increased. The model becomes simpler and is simpler to interpret as a result. If the right subset is selected, the model's accuracy will increase. For Numerical data, we created the model by using normalization and analysis of variance (ANOVA), which is a formula of statistical used to compare variances through the average of different groups. The result of ANOVA is the P-value. This value appears as the difference between group variance and the within-group variance, which ultimately produces a number that allows an inference that the null hypothesis is rejected or supported. If the difference between the features is significant, P-value will be larger, the null hypothesis is rejected. This model improves the importance of data, and it makes all numerical features significant variables because of the P-value of ANOVA $<$ 0.05. Table II shows the feature selection of numerical data.

For categorical data, we created the model by using normalization and the Chi-squared technique, which is a statistical test utilized to examine the variance among categorical variables randomly chosen to judge suitability among observed and expected results, the candidate variable for the feature removed when it irrelevant to the problem, this model appears all categorical features are significant variables because the P-value of Chi-squared $<$ 0.05. As a Table III.

	age	job	marital	education	default	balance	housing	loan	contact	day	...	duration	campaign	pdays	previous
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	...	261	1	-1	0
1	44	technician	single	secondary	no	29	yes	no	unknown	5	...	151	1	-1	0
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	...	76	1	-1	0
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	...	92	1	-1	0
4	33	unknown	single	unknown	no	1	no	no	unknown	5	...	198	1	-1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45206	51	technician	married	tertiary	no	825	no	no	cellular	17	...	977	3	-1	0
45207	71	retired	divorced	primary	no	1729	no	no	cellular	17	...	456	2	-1	0
45208	72	retired	married	secondary	no	5715	no	no	cellular	17	...	1127	5	184	3
45209	57	blue-collar	married	secondary	no	668	no	no	telephone	17	...	508	4	-1	0
45210	37	entrepreneur	married	secondary	no	2971	no	no	cellular	17	...	361	2	188	11

45211 rows × 21 columns

(a) Data before label encoding.

	age	job	marital	education	default	balance	housing	loan	contact	day	...	duration	campaign	pdays	previous	poutcome	y
0	58	4	1	2	0	3036	1	0	2	5	...	261	1	-1	0	3	
1	44	9	2	1	0	945	1	0	2	5	...	151	1	-1	0	3	
2	33	2	1	1	0	918	1	1	2	5	...	76	1	-1	0	3	
3	47	1	1	3	0	2420	1	0	2	5	...	92	1	-1	0	3	
4	33	11	2	3	0	917	0	0	2	5	...	198	1	-1	0	3	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45206	51	9	1	2	0	1741	0	0	0	17	...	977	3	-1	0	3	
45207	71	5	0	0	0	2639	0	0	0	17	...	456	2	-1	0	3	
45208	72	5	1	1	0	5455	0	0	0	17	...	1127	5	184	3	2	
45209	57	1	1	1	0	1584	0	0	1	17	...	508	4	-1	0	3	
45210	37	2	1	1	0	3779	0	0	0	17	...	361	2	188	11	1	

45211 rows × 21 columns

(b) Data after label encoding.

Fig. 3. Label encoding of dataset.

F. Label Encoding

In this stage, we will encode the label target class via transformation (no to 0 and yes to 1). In the classification process, the dependent variables are usually affected by ratio scale variables and qualitative factors. Hence, these classified variables should be transformed into numerical values utilizing encoding techniques due ML algorithms just accept the numerical inputs [17]. The categorical columns transformed so will get numeric values as a Fig. 3, by applying the function (Label Encoding) of categorical columns. Then, we will drop some columns, for example, “month”, due there is no annual information so it will make the data biased because some months contain more data than others.

G. Balancing Dataset

For classification model, the balance of the dataset is important for a model, to create models with higher accuracy devoid of bias. An uneven class deployment of the dataset can lead to trouble in later stages of classification and training, as the classifiers will have less data to understand the features of

TABLE III.
THE FEATURE SELECTION OF NUMRECL DATA

Features	Without Chi-squared	With Chi-squared
job	0.261755	0.00005
marital	-0.102826	0.00011
education	0.197275	-3.88
poutcome	-1.973561	-3.484

a particular class. The SMOTE technique is one of the best methods utilized to balance the dataset. SMOTE used the algorithms of the nearest neighbor to generate synthetic and new data that can be utilized to the train models. Unlike normal upsampling. In this paper, we used SMOTE, which it will make to generate points of new data of the minority classes to balance the dataset, which will be increased the likelihood of successfully learning [33]. Fig. 4 shows the dataset before and after balanced data.

H. Split the Dataset

After the dataset pre-processing and selected the features appropriately, the dataset becomes ready to make predictions with the ML algorithms. In this section, we do not use the test dataset due it has data leakage. As described in Kaggle the dataset of (test.csv) has about 4521 rows and 18 columns, it represents 10% of the dataset train (train.csv) randomly selected, this is very harmful to the models since it gives known data during the model fitting process and gives unrealistic results as a consequence. Therefore, we used the function (pd. concat) to connect two datasets of the train data size (45211, 17), test data size (4521, 17) into one dataset with size (45932, 17), as Fig. 5, and split the dataset to (training=0.8, testing=0.2) k–fold cross validation (k=5).

I. Applying Models and Results

The dataset was read into the Jupyter notebook with the use of Python code after the data preprocessing was finished, and the findings were then reviewed with some explanation. The imported data was scanned for missing values using the isnull () sum () function, and outliers from the data set were identified and will be appropriately addressed in the modeling stage. The data set employed for this study project is also reliable. Has the customer subscribed to a term deposit? is the target variable, which is a binomial classification problem that can only be answered with a no or yes based on the initial data set. Yes is encoded to 1 and no to 0 since the selected algorithms perform best with scalar values. Categorical variables are transformed into scalar values using the factor function. Following the completion of all these actions. finally, splitting the dataset and applying ML algorithms to predictions.

J. Evaluation

Four frequently utilized algorithms are employed to assist in finding patterns in the acquired data once testing has been completed and the trained model has been loaded, as was previously described. At this point of the investigation, the outcomes produced by the algorithms will be shown and discussed. To understand the misclassification errors regarding each algorithm or approach by which each value was measured for accuracy, the findings of the analysis were computed for this study depending on the confusion matrix. Table IV indicates the results of these models. Fig. 6 shows the ROC curves for the classifier models. Fig. 7 displays the execution time for models.

VI. CONCLUSION

In this section, the performance results are displayed and discussed in terms of accuracy, F1 score, recall, precision, AUC, and ROC curves. The proposed methodology includes preprocessing of the dataset, data normalization with ANOVA and chi-square to determine the best features and improvement, balanced the dataset to adequate with algorithms, two datasets are connected into one dataset to obtain accurate predictions. The algorithms used as (Random Forest Classifier, Decision Tree Classifier, K_Neighbors classifier and CatBoost classifier were the accuracy results (0.97, 0. 95, 0. 91, 0. 98) respectively. And execution time (7.98 sec, 0.4 sec., 36.09 sec., 15.04 sec.), respectively. Although the execution time of CatBoost was 15.04 of seconds, it gave the highest f1_score and accuracy. Finally, the study's future goals, the proposed methodology involves being created a web page, thereby helping many banking institutions in the speed and accuracy of prognosis to access if the product (bank term deposit) by the customer would be (yes) or (no). Using more techniques of feature selection in conjunction with the marketing dataset to improve diagnosis. In addition, using deep learning models will improve the prediction of bank deposits.

ACKNOWLEDGMENT

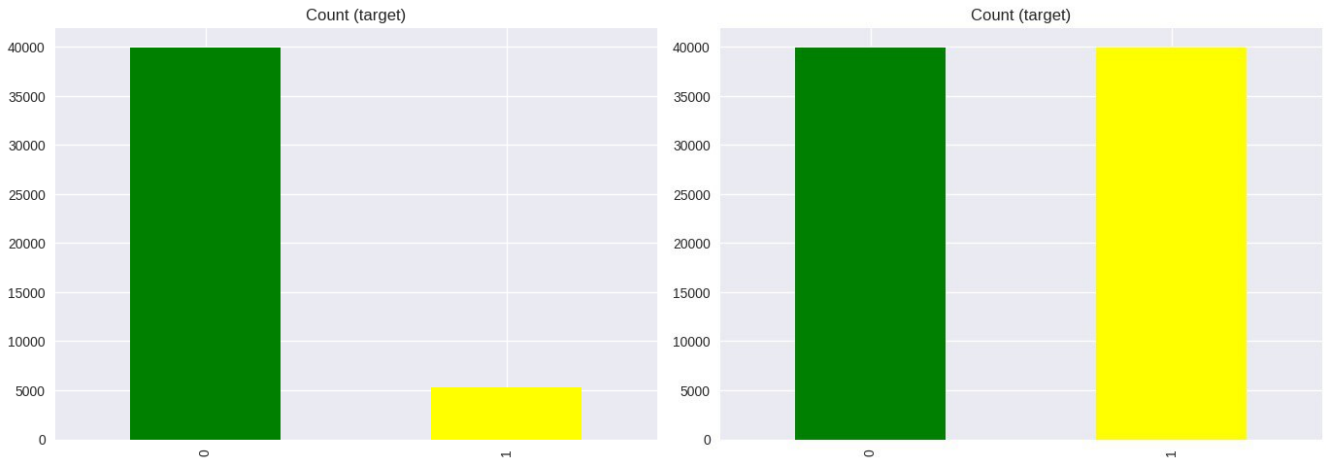
We Introduce our thanks to Iraqi Journal for Electrical and Electronic Engineering for helping the students and the lecturers.

CONFLICT OF INTEREST

The authors have no conflict of relevant interest to this article.

TABLE IV.
THE RESULT OF MODELS

Models	Confusion Matrix		Training Accuracy	Testing Accuracy	F1_ Score	Recall	Precision
	TP	FP					
	FN	TN					
Random Forest	1132	250	0.99	0.97	0.90	0.99	0.82
	12	7618					
Decision Tree	1126	400	0.96	0.95	0.84	0.98	0.74
	18	7468					
KNN	1114	800	0.92	0.91	0.73	0.97	0.58
	30	7068					
CatBoost	1138	219	0.99	0.98	0.91	0.99	0.84
	6	7649					



(a) Imbalanced dataset (b) Balanced data

Fig. 4. Balanced the data with SMOTE technique.

REFERENCES

[1] Z. Ruan and K. Siau, "Digital marketing in the artificial intelligence and machine learning age," in *Americas Conference on Information Systems (Amcis 2019)*, (Cancun, Mexico), 2019.

[2] S. Dimitrieska, A. Stankovska, T. Efremova, *et al.*, "Artificial intelligence and marketing," *Entrepreneurship*, vol. 6, no. 2, pp. 298–304, 2018.

[3] T. Ribeiro and J. L. Reis, "Artificial intelligence applied to digital marketing," in *Trends and Innovations in Information Systems and Technologies: Volume 2 8*, pp. 158–169, Springer, 2020.

[4] T. Ribeiro and J. L. Reis, "Artificial intelligence applied to digital marketing," in *Trends and Innovations in Information Systems and Technologies* (Á. Rocha, H. Adeli, L. P. Reis, S. Costanzo, I. Orovic, and F. Moreira, eds.), (Cham), pp. 158–169, Springer International Publishing, 2020.

[5] M. S. Ullal, I. T. Hawaldar, R. Soni, and M. Nadeem, "The role of machine learning in digital marketing," *Sage Open*, vol. 11, no. 4, pp. 1–12, 2021.

[6] K. Bayoude, Y. Ouassit, S. Ardchir, and M. Azouazi, "How machine learning potentials are transforming the practice of digital marketing: state of the art," *Periodicals of Engineering and Natural Sciences*, vol. 6, no. 2, pp. 373–379, 2018.

[7] A. S. Rosokhata, O. I. Rybina, A. O. Derykolenko, and V. Makerska, "Improving the classification of digital marketing tools for the industrial goods promotion in the globalization context," *Research in World Economy*, vol. 11, no. 4, pp. 42–52, 2020.

	age	job	marital	education	default	balance	housing	loan	contact	day	...	duration	campaign	pdays	previous	poutcome	y
0	58	4	1	2	0	3036	1	0	2	5	...	261	1	-1	0	3	0
1	44	9	2	1	0	945	1	0	2	5	...	151	1	-1	0	3	0
2	33	2	1	1	0	918	1	1	2	5	...	76	1	-1	0	3	0
3	47	1	1	3	0	2420	1	0	2	5	...	92	1	-1	0	3	0
4	33	11	2	3	0	917	0	0	2	5	...	198	1	-1	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
49727	33	7	1	1	0	119	1	0	0	30	...	329	5	-1	0	3	0
49728	57	6	1	2	1	0	1	1	2	9	...	153	1	-1	0	3	0
49729	57	9	1	1	0	558	0	0	0	19	...	151	11	-1	0	3	0
49730	28	1	1	1	0	1187	0	0	0	6	...	129	4	211	3	1	0
49731	44	2	2	2	0	1186	1	1	0	3	...	345	2	249	7	1	0

49732 rows × 21 columns

Fig. 5. Display dataset after contact.

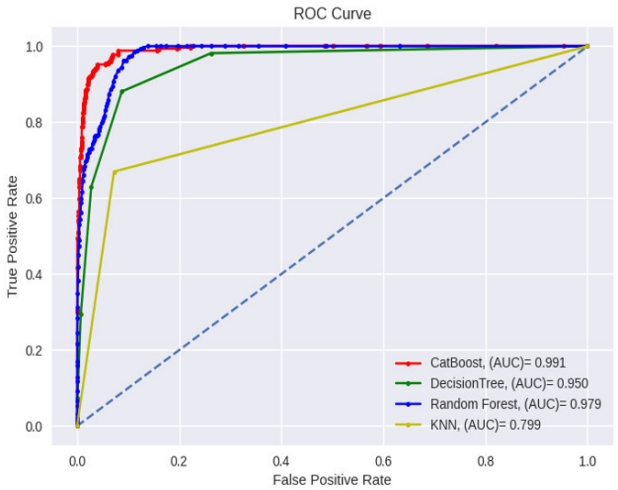


Fig. 6. Display the ROC curve for models.

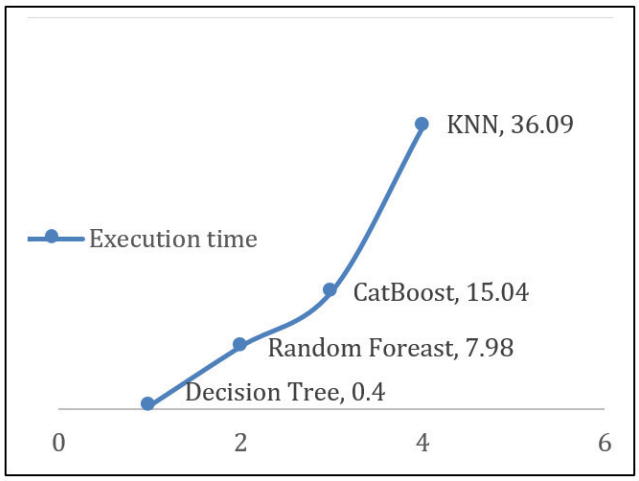


Fig. 7. Execution time for classifier models.

[8] W. Gao and Z. Ding, "Construction of digital marketing recommendation model based on random forest algorithm," *Security and Communication Networks*, vol. 2022, pp. 1–9, 2022.

[9] D. C. Gkikas, P. K. Theodoridis, and G. N. Beligiannis, "Enhanced marketing decision making for consumer behaviour classification using binary decision trees and a genetic algorithm wrapper," in *Informatics*, vol. 9, p. 45, MDPI, 2022.

[10] Z. Li, "Accurate digital marketing communication based on intelligent data analysis," *Scientific Programming*, vol. 2022, pp. 1–10, 2022.

[11] A. De Mauro, A. Sestino, and A. Bacconi, "Machine learning and artificial intelligence use in marketing: a general taxonomy," *Italian Journal of Marketing*, vol. 2022, no. 4, pp. 439–457, 2022.

[12] J. J. R. Angelina, S. Subhashini, S. H. Baba, P. D. K. Reddy, P. S. K. Reddy, and K. S. Khan, "A machine learning model for customer churn prediction using catboost classifier," in *7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, (Madurai, India), pp. 166–172, IEEE, 2023.

[13] M. M. Prabhakaran and J. Anandhi, "Artificial intelligence applied to digital marketing in machine learning,"

Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 10, no. 3, pp. 1035–1042, 2019.

[14] N. A. Noori and A. A. Yassin, "Towards for designing intelligent health care system based on machine learning.," *Iraqi Journal for Electrical & Electronic Engineering*, vol. 17, no. 2, 2021.

[15] A. Kaponis and M. Maragoudakis, "Data analysis in digital marketing using machine learning and artificial intelligence techniques, ethical and legal dimensions, state of the art.," in *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, (Corfu, Greece), pp. 1–9, 2022.

[16] V. Mitić *et al.*, "Benefits of artificial intelligence and machine learning in marketing," in *Sinteza 2019-International scientific conference on information technology and data related research*, (Belgrade), pp. 472–477, Singidunum University, 2019.

[17] J.-A. Choi and K. Lim, "Identifying machine learning techniques for classification of target advertising," *ICT Express*, vol. 6, no. 3, pp. 175–180, 2020.

[18] L. Ma and B. Sun, "Machine learning and ai in marketing–connecting computing power to human insights," *International Journal of Research in Marketing*, vol. 37, no. 3, pp. 481–504, 2020.

[19] Y. Zheng, "Decision tree algorithm for precision marketing via network channel," *Computer systems science and engineering*, vol. 35, no. 4, pp. 293–298, 2020.

[20] S. Armas-Arias, C. Páez-Quinde, L. Ballesteros-López, and S. López-Pérez, "Decision trees for the analysis of digital marketing in the tourism industry: Tungurahua case study," in *Emerging Research in Intelligent Systems* (M. Botto-Tobar, H. Cruz, A. Díaz Cadena, and B. Durakovic, eds.), (Cham), pp. 351–361, Springer, 2021.

[21] K. Arora, M. Faisal, *et al.*, "The use of data science in digital marketing techniques: Work programs, performance sequences and methods.," *Startupreneur Business Digital (SABDA Journal)*, vol. 1, no. 2, pp. 143–155, 2022.

[22] A. Miklosik, M. Kuchta, N. Evans, and S. Zak, "Towards the adoption of machine learning-based analytical tools in digital marketing," *Ieee Access*, vol. 7, pp. 85705–85718, 2019.

[23] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of k-nearest neighbour (knn) algorithm and its different variants for disease prediction," *Scientific Reports*, vol. 12, no. 1, p. 6256, 2022.

[24] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *KNN Model-Based Approach in Classification* (R. Meersman, Z. Tari, and D. C. Schmidt, eds.), (Catania, Italy), pp. 986–996, Springer, 2003.

[25] O. S. Atiyah and S. H. Thalij, "A comparison of coivd-19 cases classification based on machine learning approaches," *Iraqi Journal for Electrical and Electronic Engineering*, vol. 18, no. 1, pp. 139–143, 2022.

[26] B. Yang and J. Li, "Precise marketing strategy optimization of e-commerce platform based on knn clustering," *Journal of Mathematics*, vol. 2022, pp. 1–8, 2022.

[27] J. T. Hancock and T. M. Khoshgoftaar, "Catboost for big data: an interdisciplinary review," *Journal of big data*, vol. 7, no. 1, pp. 1–45, 2020.

[28] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the catboost classifier with other machine learning methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, pp. 738–748, 2020.

[29] R. Sanjeetha, A. Raj, K. Saivenu, M. I. Ahmed, B. Sathvik, and A. Kanavalli, "Detection and mitigation of botnet based ddos attacks using catboost machine learning algorithm in sdn environment," *International Journal of Advanced Technology and Engineering Exploration*, vol. 8, no. 76, pp. 445–461, 2021.

[30] P. Rathi, "Banking dataset - marketing targets," 2021. https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets/code.

[31] R. Rodríguez-Pérez and J. Bajorath, "Feature importance correlation from machine learning indicates functional relationships between proteins and similar compound binding characteristics," *Scientific reports*, vol. 11, no. 1, p. 14245, 2021.

[32] J. R. Saura, "Using data sciences in digital marketing: Framework, methods, and performance metrics," *Journal of Innovation & Knowledge*, vol. 6, no. 2, pp. 92–102, 2021.

[33] K. Li, W. Zhang, Q. Lu, and X. Fang, "An improved smote imbalanced data classification method based on support degree," in *International conference on identification, information and knowledge in the internet of things*, (Beijing, China), pp. 34–38, IEEE, 2014.