Identifying Long-Term Deposit Customers: A Machine Learning Approach

Mohammad Abu Tareq Rony [1], Md. Mehedi Hassan [2], Eshtiak Ahmed [3]
Asif Karim*[4], Sami Azam [5], D. S. A. Aashiqur Reza [6]
Department of Statistics, Noakhali Science and Technology University, Noakhali, Bangladesh[1]
Department of Computer Science and Engineering, North Western University, Khulna, Bangladesh[2]
Faculty of Infomration Technology and Communication Sciences, Tampere University, Tampere, Finland [3]
College of Engineering, IT and Environment, Charles Darwin University, Darwin, NT, Australia[4,5]
Mathematics Discipline, Khulna University, Khulna, Bangladesh[6]
Emails: abutareqrony@gmail.com[1] mehedihassan@ieee.org[2] (ORCID: 0000-0002-9890-0968)[2]eshtiak.ahmed@tuni.fi[3], asif.karim@cdu.edu.au[4], sami.azam@cdu.edu.au[5], and aashiq.reza007@gmail.com[6]

***Abstract*—Majority of the revenue from the banking sector is usually generated from long term deposits by customers. It's very important for banks to understand customer characteristics to increase product sales. To aid this, marketing strategies are employed to target potential customers and let them interact with the banks directly, generating a big amount of data on customer characteristics and demographics. In recent years, it has been discovered using various data analysis, feature selection, and machine learning techniques can be employed to analyze customer characteristics as well as variables that can impact customer decision significantly. These methods can be used to identify consumers in different categories to predict whether a customer would subscribe to a long-term deposit, allowing the marketing strategy to be more successful. In this study, we have taken a R programming approach to analyze financial transaction data to gain insight into how business processes can be improved using data mining techniques to find interesting trends and make more data-driven decisions. We have used statistical analysis like Exploratory Data Analysis (EDA), Principal Component Analysis (PCA), Factor Analysis and Correlations in the given data set. Besides, the study's goal is to use at least three typical classification algorithms among Logistic Regression, Random Forest, Support Vector Machine and K-nearest neighbors, and then make predictive models around customers signing up for long term deposits. Where we have gotten best accuracy from Logistic Regression which is 90.64% as well the sensitivity is 99.05%. Results were analyzed using the accuracy, sensitivity, and specificity score of these algorithms.**

***Index Terms*—Telemarketing, Classification, PCA, Machine learning, Classification**

I. Introduction

Bank marketing campaigns are usually implemented in one of two ways. One is through mass marketing campaigns directed at the general population, and the other is through targeted marketing campaigns aimed at a small group of people. According to the same report, mass marketing campaigns have a shallow positive response rate to purchase a product or subscribe to a service compared to targeted marketing [1]. As a result, much money is wasted on mass campaigns, even though they help sell the commodity. However, the primary motivation is to sell the thing, which can be accomplished effectively using direct marketing [2]. For example, indirect telemarketing, a salesperson calls a customer on the phone or cell phone to sell a product [2]. However, finding potential customers from a specific group of people is difficult. Because of the emergence of data-driven decisions in recent years, marketing managers have begun to employ statistical strategies to identify potential buyers for a product [3]. This helps them identify which customers are more likely to invest in the bank, avoiding potential impasses in the process.

Due to the recent economic turmoil in many countries, banks need to sell more long-term deposits to raise their financial reserves [4]. As a result, marketing managers are under pressure to sell long-term stakes to the general public [4].It's imperative for marketing managers to use their scarce resources to make fewer calls to clients while selling more, i.e., raise the positive response rate. With a view to achieving that, bank marketing managers may use multivariate data classification methods to identify potential customers as they already possess data from previous campaigns to analyze [5].

II. Background

As scientific computing has become more available, data-driven approaches to finding potential customers have become increasingly common in recent years [6]. Previous works on the bank telemarketing data sets have been solely based on making class predictions using various classifying algorithms [6]. Usage of the Decision Tree Algorithm and Rough Set Theory has been discussed in a study [6] where attempted to define guidelines that would be useful in determining the likelihood of a consumer signing up for a term deposit. Here only three attributes were used in the Rough Set Theory approach to make class predictions. In analyzing the bank marketing data set [7], the most essential attributes were Age, Balance, and Length. The accuracy selection for the decision tree technique used was based on the gain ratio provided for all attributes. The role of neural networks to make class predictions were also looked at for the bank marketing data set [7]. The findings revealed that neural networks could learn patterns that can predict when customers sign up for a term deposit. Aside from that, one paper used the rminer(add ref

Preprint
IEEE International Conference on Informatics and Software Engineering, 2021, Turkey

for rminer library) library to perform research on the data set. The classifiers used were the decision tree, Support Vector Machines (SVM), and

Naive Bayes [8], the most important features used by the classifiers for making predictions were calling length and month of contact [9]. Healthcare, finance, retail, intelligence, and telecommunications are only a few of the currently used characteristics in data mining tools and techniques to develop their operations [10]. For example, data mining is being used in the banking industry to create models for risk analysis, Customer Relationship Management (CRM), direct marketing, and credit card fraud prevention. Data mining has gained attention in the healthcare industry to prevent medical insurance fraud and violence and predict patient behavior as well as well-being patterns.

III. METHODOLOGY

This paper aims to identify the main features and develop a Machine Learning model that can divide a client list into two groups: those who want to subscribe to a long-term deposit and those who do not. Due to the difficulty of high dimensional in the dataset that typically accompanies an imbalanced dataset, sampling techniques, and algorithm-based approaches may not resolve the class imbalance. To solve this problem, a feature selection approach (PCA, Factor analysis, etc.) is required to select a subset of features that will help in the model's optimal efficiency. When operating with a dataset with high dimensionality, the feature selection process is essential. PCA is a system that reduces the dataset's dimensions, improves interpretability, and does not lose information. It creates new uncorrelated variables in a step-by-step process to reduce variance. We have employed factor analysis to find associations and then Five ML classifiers to predict output class. Besides, this study aims to build a reliable and realistic recommendation algorithm for predicting client take-up based on the client type. The target value is a binary class yes or no regarding the clients' term deposit subscription. The task could be solved using logistic regression, Random Forest, SVM, KNN, and MLP to classify the data. We plan to use these algorithms and compare their accuracy to determine which algorithm produced the most accurate results and find the best method for producing a classifier with better predictive capabilities that could be used to decide whether a customers would sign up for a term deposit. Correlation analysis was used to help determine accurate rules that banks could use to identify customers who were likely to subscribe to a term deposit. The discovered practices could aid in deciding which subgroups within a class are more likely to subscribe. This may be due to their age, marital status, gender, educational status, or other factors. This involves deciding if the personal efforts put forward by banks in their campaign process affect how well consumers will subscribe to this service. For instance, the number of phones calls each consumer has received, the results of previous campaign exercises, and so on. These characteristics were looked at to see if any patterns emerge which could improve marketing campaigns and services.

A. Dataset Overview

The UCI Machine Learning Repository provided a dataset on bank telemarketing for this project. A Portuguese bank used its call center to execute direct marketing campaigns to encourage and draw customers to their term deposit scheme to develop their business. The dataset contains 17 campaigns that ran between May 2008 and November 2010. A compelling long-term deposit application with fair interest rates was made available on the internet during the campaign. The progress rate is 11.27%, according to the findings. Consumer characteristics and other factors were discovered to affect the response [11].

B. Data Pre-processing

To increase data quality, the primary dataset was preprocessed. Preprocessing techniques were applied to several attributes. Furthermore, some continuous details were transformed to categorical attributes because specific algorithms, such as logistic regression can only be used to categorical attributes. Moreover, features with a large number of categories were merged into a single attribute. Since the dataset was small enough to accommodate, the sampling approach was not used. Therefore, the results apply to the whole dataset. Dimensionality reduction was needed because the number of attributes was significantly more than the number of data instances. All data instances with unknown values in a class were imputed, and all unknown values for features other than the class were assigned by the most frequently occurring value. Finally, the dataset needed to be formatted so that the RStudio data analysis tool could understand. The age ranges for the Age Group attribute in the original dataset ranged from 17 to 98. Simply converting each age value into a single category resulted in 78 categories, some of which have only one case. Since specific types are vastly underrepresented, it is not easy to assess how the attribute affects the test class when there are many categories. We don't use output class in our PCA review because we can only use PCA for unsupervised ML. Furthermore, since PCA cannot deal with character values, we excluded them while planning our data for PCA analysis.

IV. EXPLORATORY DATA ANALYSIS

For data visualization techniques, exploratory data analysis (EDA) has been used to analyze and examine data sets and summarize their key features. It also helps determining if the statistical approaches considered for data analysis are appropriate. [9].

To begin with, it seems that banks are not particularly interested in contacting the elderly. In Fig.1 the relative frequency is higher when y = 1 after 60 years, it is lower when y = 0. To put it another way, older adults are more likely to invest in a term deposit. We can also divide the age function into three groups at 30 years: [0, 30], [30, 60], and [60, +Inf]. The minimum and maximum values are 17 and 98, respectively, but new findings are likely to occur outside of this range. This categorical variable would take the place

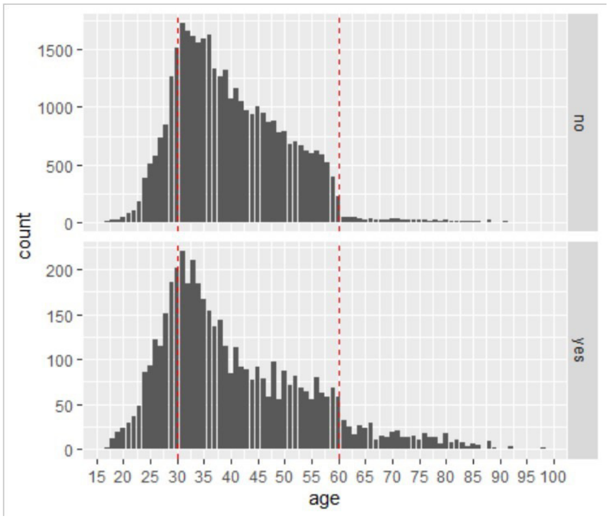

Fig. 1: Distribution of Age

of the continuous variable" age". Although the continuous to-discrete conversion may have resulted in some information loss, there was no discernible pattern between years. It would be easier to grasp the algorithms if they were divided into classes.

A. Gender

According to the dataset in Fig.2 and 3, males had a marginally better response rate, but only by a slim margin. Given the size of the dataset, it's hard to decide if it is true for every scenario. There is always a possibility of having more female respondents with more data.

B. Age Range

Younger men between the ages of 21 and 30, followed by those between the ages of 31 and 40, made up most respondents. The response rate was lowest for those over the age of 41. This may be because older people, rather than younger people, chose not to participate in the survey or were never permitted to in Fig.2 and 3.

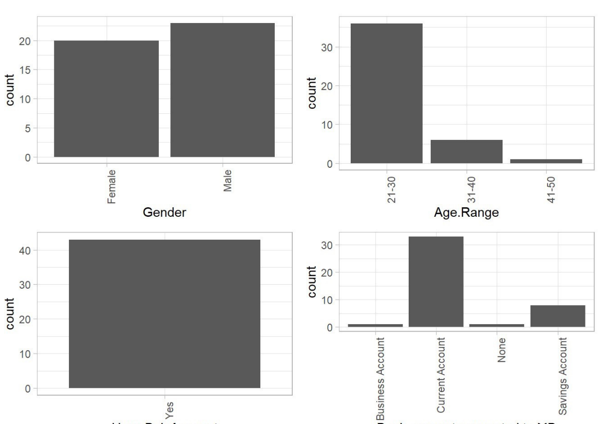

Fig. 2: Gender and Age Range

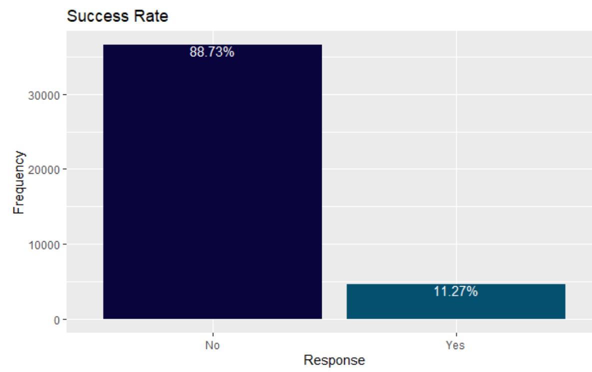

Fig. 3: Success Rate

V. Factor Analysis

Factor analysis reduces the number of variables in a dataset to a manageable number [22]. This method converts the highest common variance of all variables into a single rating. For future research, we may use this score as an index of all variables. We used factor analysis to find essential variables from a broad dataset in our dataset [13]. For factor analysis, we take some variables which are converted as a dummy variable first and preprocess for factor analysis. Factor Analysis with system = minres and fa () feature applied to 6 factors since it found that six components provide the most variance, which is why the factor analysis model is fitted with six elements. It's worth noting that the two strategies yield similar outcomes, but they present the data differently and present the variables in different ways. The performance produced using the minres method; could have also chosen age, period, campaign, etc. It starts with a matrix that displays the three variables (MR2, MR3, and MR1) and a few other figures. This can be obtained by exercising. pc$loadings is a variable in pc$loadings. Only the 'important values' will be displayed–values similar to 0 will be hidden for more straightforward analysis. Here it can be see married dummy 0.732 is a good variance for MR1.h2 is the commonality score–a sum of the squared factor loadings for each question. 1-h2 is the individuality score, and u2 is the uniqueness score. The variation that is 'unique to the variable and not associated with other variables is referred to as uniqueness. It's the same as 1 – communality (variance that is shared with other variables). The higher the 'uniqueness,' the lower the variable's relevance in the factor model. E.g., the age variable has a uniqueness of -0.0021, which is higher than other variables. And test the hypothesis that six factors are sufficient. Following that, several model-related figures are shown. It then gives a chi-squared test to see whether the model is complex enough (or whether more parameters are needed) and other diagnostics to see whether the model is suitable. It'll try five using the ML optimization approach since we think there maybe five. Besides, it can see the best connections between MR and variables on the map.

VI. PRINCIPAL COMPONENT ANALYSIS ALGORITHM

PCA is a dimensionality-reduction method for reducing the dimensionality of big data sets by translating a large group of variables into a smaller one that preserves the maximum of the information in the large group. We can see from our data that most of the variance is in the first principle variable [14].The principle components reflect the data directions that explain the most variation and the lines that encapsulate the most data information. We consider age, duration, campaign, pdays, previous, cons.price.idx, cons.conf.idx, euribor3m, nr.employed for PCA. PCA can do only for numerical and integer values. Hence, we can convert a character into label data, but it will bring some complexity. After preprocessing data, we see the correlations between new data. For example, age variables are positively correlated with the effect of campaign variables that are at least 0.0045 and negatively correlated with a duration that is -0.00086. The scale of the dataset for PCA is applied because all the variable was not in the same units. From a summary of the PCA, we find a standard deviation of 10 principle components with the proportion of variance and cumulative proportion find 0.91537 cumulative proportion in PC6, which is enough for measuring the variance of a new dataset. So in Fig. 6 there is a maximum standard deviation in PC1 that is 1.9737(feature Score), and it decreased gradually. From the eigenvalue, we can say the six-principle component is enough for measuring the variability of new data as it reaches more than 90% variance proportion, which is good and there is no significance in PC10. After doing all the above, we find loadings, scores of PCA analysis. We can see that there is maximum variance in component one that is PC1, and it is reducing gradually. So, we can say that there is the maximum variation in PC1 which will help analyze more accurately and indicate the more significant variance of the new dataset. From the figure, we can say there is more than 90% variance computed from 6 components in Fig.6 where summative variance, also known as multivariate variability, is referred to as component variance in y-axis.

Fig.6 allows us to see the associations between the variables of PCA. It can see that the variable categories here are days, cons.price.idx, cons. conf.idx, euribor3m, nr. employed are in PC1 with positive variance. It also allows us patterns of correlation between the variable categories. This Scree(Fig.6) diagram shows how variance is reduced concerning increasing the number of principle components. The maximum amount of component variance is in between PC1-PC6.

A. Correlation Analysis

Correlation is the linear relationship between two continuous variables. Correlation is widely used where there is no known response component [12]. Euribor 3 months rate, Employment Variation Rate, Consumer Price Index, Consumer Trust Index The number of people who work at the bank as well as Social and economic metrics make up all five continuous variables in Fig.7. They're supposed to be inextricably connected.

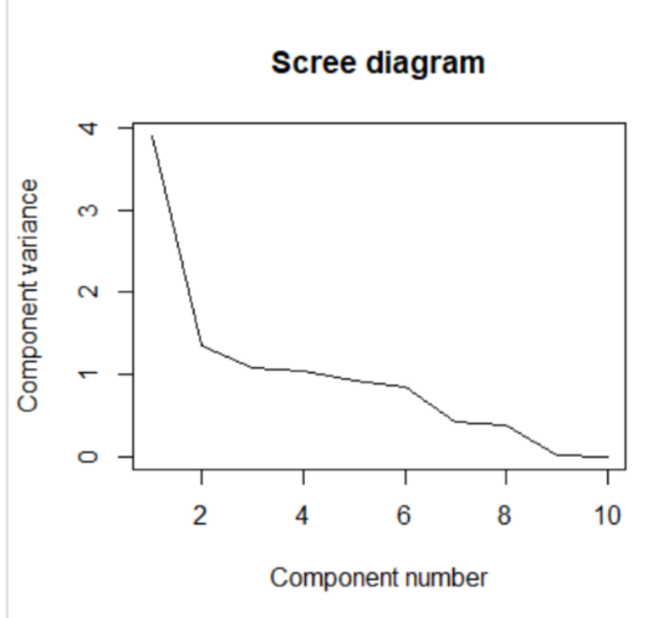

Fig. 4: Scree Diagram

Fig. 5: Correlation Analysis

A correlation coefficient of 0.90 or higher exists in three pairs. Let's see which variables can be omitted from this correlation matrix to lighten it up. The value of the emp.var.rate is unimportant. There are changing staff before each campaign would be detrimental for banks. To smooth out the similarities between our five variables, we'll remove them. Both euribor3m and nr. employed are kept because they are closely related (0.95).

VII. MACHINE LEARNING ALGORITHMS

Machine learning is a form of data analysis that allows for analytical models to be automated. It's a branch of artificial intelligence based on the idea that computers can learn from data, analyze data, and make decisions with minimal human intervention [15]. Researchers aim to use supervised learning to predict a target performance (y) [2]. To determine how to achieve the highest degree of accuracy, it's imperative to first understand the capabilities of the intended algorithm. To build a classification tool that can be used in various domains, there needs to be a prior understanding of how classification algorithms react to datasets. This study compared the results obtained using the sampling technique, algorithm-based technique, and feature selection technique for handling class accuracy predictions using five classification algorithms. These are listed below, along with a brief overview of their guiding principles and methodologies [16].

A. Logistics Regression (LR)

When the dependent variable is dichotomous, logistic regression is the safest regression analysis to use (binary). To classify data and find the relationship between one dependent binary variable and one or independent variable, logistic regression is used [17].

B. Random Forest (RF)

Random forest is a supervised learning algorithm that can classify and predict data. The randomized method of locating root nodes to break features distinguishes it from decision trees. Random forest is good at dealing with missing values. The over-fitting problem is a potential downside of this algorithm unless a sufficient number of trees are created to improve prediction accuracy. [18].

C. Support Vector Machine (SVM)

An SVM is a supervised machine learning model for two-group classification problems that employ classification algorithms. SVM models will categorize new text after being given sets of labeled training data for each group [19].

D. K nearest neighbors (KNN)

The KNN learning algorithm is another classification algorithm that indicates the simple majority of KNN as a query prediction based on the shortest distance between instance and sample. Because of its predictive power and fast calculation time, KNN is commonly used, and it typically produces highly competitive results [20].

E. Multilayer Perceptron (MLP)

An MLP is a form of artificial neural network that uses feedforward learning (ANN). MLP is a fuzzy term that can apply to any feedforward ANN, or it can refer to networks made up of several layers of perceptrons [21].

VIII. Evaluation Metrics

For executing the different machine learning models and finding the best one, various evaluation metrics are there [23]. Different evaluation techniques are introduced based on the confusion metrics such as accuracy, Sensitivity, and Specificity our model evaluation is done based on these evaluation criteria [19].

- **Accuracy:** It represents how many instances are correctly predicted.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{1}$$

- **Sensitivity:** The ratio of TN that is correctly identified is known as sensitivity.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

- **Specificity:** The ratio of TN that is correctly identified is known as specificity.

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

TABLE I: Comparison Between classification algorithm

Algorithms	Accuracy	Sensitivity	Specificity
LR	0.9064	0.9905	0.2222
RF	0.9016	0.9795	0.2667
SVM	0.8810	0.9615	0.2456
KNN	0.8991	0.9877	0.1778
MLP	0.8732	0.9455	0.1674

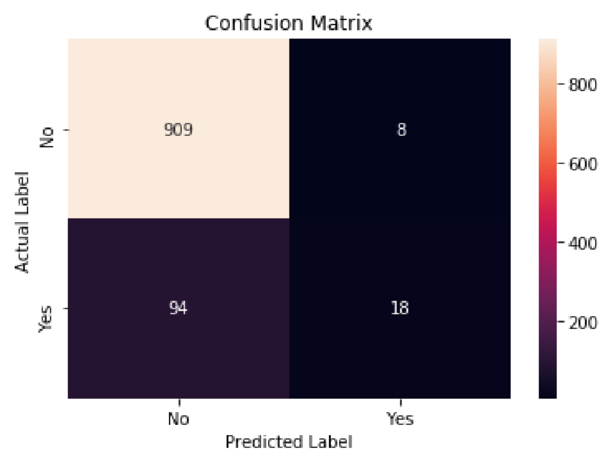

Fig. 6: Confusion Metrics(LR) on Test Set

IX. Results and discussion

We have used PCA and Factor Analysis to determine the value of each variable and a correlation test to determine how the variables are related. The logistic regression model is used for predicting whether or not a client would subscribe to a term deposit. By a fraction of a percent, the logistic model bettered the RF, SVM, KNN, and MLP models in Table 1. Because the five models' accuracies are so similar, all of them could be used in this scenario. It will rely on what additional information was sought from the study if any. If it's just about estimation and which model is the most reliable, the logistic regression model is the way to go. Since these two models are at opposite ends of the complexity continuum, it would be essential to see how they compare to each other and the other three models discussed here. Another field that should be investigated is using a variable selection approach to reduce the number of variables in these models and see if this reduces the prediction accuracy. Finally, we look at some machine learning classification models to see how effective our model is, and we see that logistics regression outperforms other classification algorithms.

On top of these, a more extensive and more diverse dataset including client-based data and real-world data can cause a significant difference in this research approach. It would also be interesting to see whether high-quality predictive models can be generated without contact-based data.

X. Conclusions and Future Work

In this article, we have used various data analysis techniques to understand potential customer behavior to aid marketing campaigns for banks. A real-world and recent dataset from a Portuguese bank has been used for this analysis. Logistic regression was proved to be the best option for creating the model excellent predictive abilities. We have evaluated the input value in the Logistic Regression model using accuracy score, sensitivity and specificity analysis, and managers may use this information to optimize campaigns (for example, by asking agents to make longer phone calls or scheduling campaigns for specific months). Due to the clarity of the derived regulations, it can potentially assist decision-makers in developing a plan for attracting and targeting customers, making faster and better loan approval decisions, and reducing management risk. As a result, rather than analyzing a data set based on attributes, and intractable algorithm is used to extract meaningful patterns. Better and more complex classification methods may be used to predict a customer's decision. Other high-performing classification models, include more features, and a more efficient prediction tool can be developed in future extension of this research as well to get more satisfactory results. There is still scope for performing a more descriptive analysis to gain more information and determine whether a factor analysis is necessary. Many different models will calculate each class's predictability and the most contributing details on some attributes.

References

[1] M. Sérgio, L. Raul and C. Paulo, "Using data mining for bank direct marketing: an application of the CRISP-DM methodology", RepositoriUM, 2011.

[2] C. S. T. Koumétio, W. Cherif and S. Hassan, "Optimizing the prediction of telemarketing target calls by a classification technique," 2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM), 2018

[3] J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T), 2017

[4] M. Sérgio, P. Cortez, and P. Rita. "A data-driven approach to predict the success of bank telemarketing." Decision Support Systems 62 (2014): 22-31.

[5] de Matos, J. Cadete, and A. Dias. "The Portuguese Central Credit Register as a key input to the analysis of financial stability… and beyond!." IFC Bulletins chapters 46 (2017).

[6] S. Palaniappan, A. Mustapha, C. Mohd Foozy and R. Atan, "Customer Profiling using Classification Approach for Bank Telemarketing" JOIV: International Journal on Informatics Visualization 1.4-2 (2017): 214-217.

[7] D. Hassan, A. Rodan, M. Salem and M. Mohammad, "Comparative Study of using Data Mining Techniques for Bank Telemarketing Data," 2019 Sixth HCT Information Technology Trends (ITT), 2019, pp. 177-181

[8] T. Yang, K. Qian, D. C. Lo, Y. Xie, Y. Shi and L. Tao, "Improve the Prediction Accuracy of Naïve Bayes Classifier with Association Rule Mining," 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 2016

[9] A. Gupta, A. Raghav and S. Srivastava, "Comparative Study of Machine Learning Algorithms for Portuguese Bank Data," 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2021.

[10] S. Manlangit, S. Azam, B. Shanmugam and A. karim, "Novel Machine Learning Approach for Analyzing Anonymous Credit Card Fraud Patterns", International Journal of Electronic Commerce Studies 10.2 (2019): 175-202.

[11] K. Kim, C. Lee, S. Jo and S. Cho, "Predicting the success of bank telemarketing using deep convolutional neural network," 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), 2015

[12] G. Sell and D. Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4794-4798

[13] B. Williams, A. Onsman and T. Brown, "Exploratory factor analysis: A five-step guide for novices", Australasian Journal of Paramedicine, vol. 8, no. 3, 2010.

[14] A. Karim, S. Azam, B. Shanmugam and K. Kannoorpatti, "An Unsupervised Approach for Content-Based Clustering of Emails Into Spam and Ham Through Multiangular Feature Formulation," in IEEE Access, vol. 9, pp. 135186-135209, 2021

[15] A. Sarmento, K. Yeo, S. Azam, A. Karim, A. Al Mamun and B. Shanmugam, "Applying Big Data Analytics in DDos Forensics: Challenges and Opportunities", Cybersecurity, Privacy and Freedom Protection in the Connected World, pp. 235-252, 2021.

[16] C. Aggarwal and C. Zhai, "A Survey of Text Classification Algorithms", Mining Text Data, pp. 163-222, 2012.

[17] M. M. Hassan, M. A. Mamun Billah, M. M. Rahman, S. Zaman, M. M. Hasan Shakil and J. H. Angon, "Early Predictive Analytics in Healthcare for Diabetes Prediction Using Machine Learning Approach," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021.

[18] M. M. Hassan, Z. J. Peya, S. Mollick, M. A. Billah, M. M. Hasan Shakil and A. U. Dulla, "Diabetes Prediction in Healthcare at Early Stage Using Machine Learning Approach," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021

[19] A. Jiménez-Cordero and S. Maldonado, "Automatic feature scaling and selection for support vector machine classification with functional data", Applied Intelligence, vol. 51, no. 1, pp. 161-184, 2020.

[20] A. Bhavani and B. Santhosh Kumar, "A Review of State Art of Text Classification Algorithms," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1484-1490

[21] R. M. AL Mgheed, "Scalable Arabic text Classification Using Machine Learning Model," 2021 12th International Conference on Information and Communication Systems (ICICS), 2021, pp. 483-485

[22] J. Wang, B. Li and H. Ji, "Factor analysis of adsorption effect of molecular sieve purification system," 2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2021

[23] P. Patel, K. Kannoorpatti, B. Shanmugam, S. Azam and K. C. Yeo, "A theoretical review of social media usage by cyber-criminals," 2017 International Conference on Computer Communication and Informatics (ICCCI), 2017