INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

PREDICTIVE ANALYTICS MODEL TO ENHANCE BANKING DECISION MAKING USING MACHINE LEARNING

1. Sherif ELsaied Elsaied Gad, 2. prof. Dr. Nashaat ELkhameesy.

[1]Ph.D. Candidate in Sadat Academy for Management Sciences (SAMS), Department of Computer Sciences and Information Systems Sadat Academy for Management Sciences, Cairo, Egypt.
Senior teacher of computer science and information systems at Al-Azhar Al-Sharif,Eldaqahlia,Egypt

[2]Professor of Computer and Information Systems Department, Sadat Academy for Management Sciences (SAMS), Cairo, Egypt.

Email: doctorsherif2021@gmail.com, wessasalsol@gmail.com

Abstract

The present global economic crisis makes it difficult for banks to attract customers. Therefore, marketing is seen to be a useful technique for the banking industry to get clients interested in a term deposit. In banks, telemarketing is a commonly used kind of direct marketing. Customers seldom react favorably, therefore data prediction models may assist in identifying the most probable potential clients. Data mining helps direct marketing efforts succeed by foretelling which leads will sign up for term deposits. In this study, we used machine learning to the benchmark dataset of banking institutions' direct marketing campaigns to create an accurate classifier to forecast which consumer would accept a long-term deposit offer. Our research reveals the remarkable influence that machine learning methods may have on the outcome of a telemarketing campaign. Data preparation and model assessment are the two main phases. In the first phase, data must be cleaned by removing duplicate records and determining if missing values should be kept or removed, data visualization, and utilizing the response coding approach to encode category characteristics using label and one-hot encoding. The dataset is originally split into training and testing but the dataset is unbalanced so we needed to consider that while training so we used the balanced class weight approach and 10-fold cross-validation to solve the imbalanced class problem. The Random Forest algorithm is used for training and testing and a perfect classifier is achieved. The proposed system outperformed all the state-of-the-art techniques and achieved perfect classification.

Keywords: Machine Learning; Bank Deposit Prediction; Random Forest;
Bank Telemarketing; One-Hot Encoding

1. Introduction

In any economic system, [13] banks play a key role in developing and carrying out financial policies. Customers deposit money with banks to benefit from the yearly interest rates that the banks give. These deposits [14] are then utilized to make a range of loans available to a diverse group of consumers. The difference between the two interest rates is what gives the banks their profit. By providing a service that allows consumers to deposit money, banks have a big impact on the economy. These company loans and investments are essential for fostering economic growth, which is an indication of a nation's advancement. Bank deposits effectively contribute to the aggregate financing structure of economic development as a result. Additionally, the customer database has various elements that have a significant impact on a depositor's potential. However, it is not simple for a human specialist to go through the client information and discover a reliable depositor. The aim may be readily attained if the outstanding input characteristics can be identified.

Through phone calls, a salesperson may solicit potential clients' readiness to buy goods or services using the direct marketing tactic known as telemarketing [16]. Notably, several suppliers of financial services have embraced telemarketing tactics to attract new clients while improving services for current clients and catering to their requirements [6]. Direct marketing is not always successful since consumers may favor well-known institutions. It is now simple to create a range of reports via marketing campaigns as well as other sorts of information required for businesses thanks to the development of telemarketing through computer technology/mobile. Term deposits, recurring deposits, fixed deposits, deposits in savings accounts and current accounts, and many more savings plans are available from banks [15].

Machine learning now refers to modifications made to systems that perform tasks related to artificial intelligence (AI) [2]. These include things like forecasting, robot control, planning, analysis, and recognition. It investigates the analysis and development of an algorithm that can forecast data. With its customized tuning settings, machine learning is utilized to create programs. Consequently, by reacting to earlier data, they will perform more effectively. A fast-evolving method that mimics the functioning of the human mind is machine learning. It successfully resolves the selectivity-invariance conundrum and represents multi-level data [3].

In many facets of life, particularly in finance, machine learning is applied [4]. Many banks use telemarketing as their primary method of consumer interaction. Banks can determine if a certain consumer is reachable or not for direct marketing, and a machine learning approach may automate the whole feature validation procedure. The drawback of this approach is that it stresses varied weights for each component, however, in reality, bank telemarketing performance may sometimes be determined by a single, powerful factor alone, which is not achievable with this method.

Several data mining approaches, including the One-R Algorithm, the Naïve Bayes (NB) classifier, categorization, and association rule mining, have been employed in the bank direct marketing arena to categorize marketing services [1]. Data mining techniques will be used to categorize the bank customers' data after exploratory data analysis on variables has been performed to determine the link between the variables and the class variable, the relationship between two variables according to the class variable, and so forth. The classification's objective is to foretell whether a customer will sign up for a term deposit (variable outcome). To utilize the model to forecast the class of objects whose class label is unknown, classification is the process of identifying a model (or function) that explains and separates data classes or ideas. Based on the study of a set of training data, the resulting model was created (data objects whose class label is known).

To help banks come up with potential solutions for this tactic, the study's goal is to identify and assess the success predictions for bank telemarketing. Our analysis uses the bank marketing dataset from a Portuguese retail bank, which contains 45,211 phone contacts with 16 input characteristics and one decision attribute [6]. This study's approach may be seen as a road map for the reader to follow the steps performed and to utilize a strategy to find the root causes of many other issues. This essay intends to provide a fast, straightforward, and instantaneous method of selecting worthy clients. It may provide certain benefits to the bank. The bank telemarketing prediction system can determine the weight of each factor that contributes to the processing's success automatically. The same attributes are processed concerning their corresponding weight on fresh test data. A deadline might be established for the applicant to determine whether or not his or her successful bank telemarketing campaign qualifies for sanction. The success of the bank's telemarketing prediction system enables quick access to a single application for priority inspection.

In this study, we used machine learning to the benchmark dataset of financial institutions' direct marketing campaigns to create an accurate classifier to forecast which consumer would accept a long-term deposit offer. Our research reveals the remarkable influence that machine learning methods may have on the outcome of a telemarketing campaign. Data preparation and model assessment are the two main phases. In the first phase, data must be cleaned by removing duplicate records and determining if missing values should be kept or removed. Data must also be visualized before utilizing the response coding approach to encode category characteristics using label and one-hot encoding. The dataset was initially divided into training and testing, but since the dataset was uneven, we wanted to take it into account when training. To address the imbalanced class issue, we employed the balanced class weight strategy and 10-fold cross-validation. A perfect classifier is created using the Random Forest method for both training and testing. The suggested approach obtained perfect classification and outperformed all state-of-the-art methods.

The rest of the paper is organized as follows; related work explaining studies that used the same dataset is presented in the second section. The third section explains the proposed system methodology including, the dataset and the procedure of preprocessing analysis and visualization that has been done on the dataset. Also, the third section explains how the dataset is encoded and ML is used for prediction and presents the results achieved with the proposed system. Finally, we conclude the paper in the fourth section.

2. Related Work

Numerous research looked at the use of different machine learning approaches to forecasting the performance of bank telemarketing for the sale of long-term deposits. The following are recent studies that used the bank marketing dataset.

Patwary et al. [5] examined the effectiveness of ensemble learning algorithms, a unique method for predicting whether a prospective client will hold a term deposit or not. The performance of the three most popular classification algorithms, Support Vector Machine (SVM), Neural Network (NN), and NB, was examined using the bank marketing dataset. The potential for ensemble approaches to enhance the performance of fundamental classification algorithms is then explored and empirically shown. The results of the experiments show that Bagging has better performance metrics than other ensemble approaches. It has 96.62 % accuracy, 97.14 % sensitivity, and 99.08 % specificity.

Safarkhani and Moro [7] created an accurate classifier to forecast which customers will accept a bank's offer of a long-term deposit. They employed the bank marketing dataset and concentrated on resampling to decrease the unbalanced data, feature selection to reduce the complexity of data computing, and dimension reduction to reduce inefficient data modeling. The operation that was carried out showed an improvement in the classification algorithm's accuracy performance. According to the experimental findings, the J48 decision tree has a prediction accuracy rate of 94.39 %, with a 0.975 sensitivity and 0.709 specificity.

Subramanian et al. [8] used NB to effectively analyze consumer data. The conditional independence assumption underlies the effectiveness of NB, and its failure leads to inaccurate predictions. However, the NB assumption is broken in the majority of real-time consumer datasets because of connected, pointless, and noisy factors. Multi-Stage Variable Selection (MSVS) is suggested to choose the pertinent factors from the customer dataset, which helps to predicate the customer patterns intelligently, to enhance NB prediction with these client consumers. The appropriate variable subset from the customer datasets is chosen in two phases using their suggested method. The NB method is used to test a further variable subset produced from the suggested MSVS technique, and the outcomes are compared using the wrapper and filter approaches. According to the findings, the MSVS strategy outperforms the wrapper and filter approaches in terms of variable subset selection and NB prediction accuracy. Additionally, compared to wrapper and filter techniques, the suggested strategy is more computationally efficient and effective in terms of time.

Dutta et al. [9] enhanced the possibility that clients will sign up for term deposits. Term deposit subscription prospects may be influenced by bank promotional efforts and customer

Vol.10 Issue 7, Pg.: 1-20
July 2022

detail research. Convolutional layers and Recurrent Neural Network (RNN) layers are stacked as part of a deep learning-based hybrid model used in an automated system that attempts to forecast potential investments in term deposits in advance. Gated Recurrent Units (GRU) are used with RNN. Their suggested prediction model is then contrasted with various benchmark classifiers such as the multi-layer perceptron classifier, decision tree classifier, and k-Nearest Neighbor (k-NN) (MLP). The results of the experimental investigation show that the suggested model achieves an accuracy of 89.59 % and a mean square error of 0.104.

Borugadda et al. [10] examined the demand for the use of telemarketing techniques to encourage long-term bank deposits among prospective bank clients. Using a variety of machine learning methods, including Random Forest (RF), SVM, Gaussian Naive Bayes (GNB), DT, and Logistic Regression, the research examined the demand for long-term bank deposits (LR). The dataset for bank marketing is taken into account for analysis. The results show that the LR model has an accuracy of 92.48 %. The study's findings also provide banks with useful data for telemarketing policy choices on the success of bank deposits to their current and potential bank clients.

Choi and Choi [11] used the bank marketing dataset to design a machine learning method. According to the study, the following factors have no bearing on the performance of bank telemarketing: employment, marital status, education, default, housing, contact, month, and prior. However, age, balance, loan, day, length, campaign, and poutcome have an impact on these factors. Second, the accuracy rate for the whole model is 0.784, indicating that the error rate is 0.216. The accuracy that would not have the success of bank telemarketing was 75.63 percent among patients, and the accuracy that had the success of bank telemarketing was 82.61 % among clients who expected to have the success of bank telemarketing.

Phungand Khuat [12] showed how the outcome of the telemarketing campaign may be affected by machine learning approaches. Data preparation and model assessment are the two main phases. Data cleaning is the first phase involves eliminating duplicate records, determining if missing values should be removed, visualizing the data to determine whether the data set is balanced, and using the response coding approach to encode category information with the use of Laplace smoothing. Additionally, introducing a "duration" option has a significant impact on the outcome. The best classifier model was picked in the second stage using common efficient methods including KNN, LR, Linear SVM, and Extreme Gradient Boosting (XGBoost). The Area under the Receiver Operating Characteristic score is used to assess the effectiveness of the study because of the bias in the data set. The most accurate approach is KNN, which performs at 91.07 percent accuracy and has a 93 percent AUC. The experiment's findings demonstrate that the best KNN with k larger than five may help understand the viewpoint from a commercial perspective.

3. Methodology

In this section, the proposed system approach is explained, along with the dataset and the steps taken for preprocessing, analysis, and visualization of the dataset. Additionally, it describes how the dataset is encoded, how machine learning is utilized for prediction, and it presents the outcomes obtained with the proposed methodology.

3.1.Dataset

The data of Portugal bank marketing efforts are taken from Kaggle [6], which is a collection of 45211 records, each with 17 qualities, to achieve the study's goal. The qualities suggest the relevant elements that have an impact on campaign outcomes. The goal variable indicates whether or not a consumer makes a term deposit. In this study, a binary classification issue is therefore handled. Table 1 offer. Information on the characteristics in the dataset, including their kinds and purposes. The dataset's term deposit subscription tendency is shown in Fig1 as a distribution. The dependent variable used in the classification process is the attribute y.

Table 1: Dataset Characteristics

Attribute Name	Type of attribute	Description	Values Present
Age	Numeric	Customer's Age	18-95
Job	Categorical	Customer's Profession	'management', 'technician', 'entrepreneur', 'blue', 'unknown', 'retired', 'admin', 'services', 'self', 'unemployed', 'housemaid', 'student'
Marital	Categorical	Marital Status of Customer	'married', 'single', 'divorced'
education	Categorical	Education Qualification	'tertiary', 'secondary', 'unknown', 'primary'
Default	Categorical	Whether the customer has credit in default	'no', 'yes'
Balance	Numeric	Balance present in account	-8019-102127
Housing	Categorical	Whether the customer has housing loan or not	'yes', 'no'
Loan	Categorical	Whether the customer acquireshousing loan or not	'no', 'yes'
Contact	Categorical	contact communication type	'unknown', 'cellular', 'telephone'
Day	Numeric	last contact day	1-31
Month	Categorical	last contact month of year	'may', 'jun', 'jul', 'aug', 'oct', 'nov', 'dec', 'jan', 'feb', 'mar', 'apr', 'sep'
Duration	Numeric	last contact duration,	0-4918

		in seconds	
Campaign	Numeric	number of contacts performed during this campaign and for this client	1-63
Pdays	Numeric	number of days that passed by after the client was last contacted from a previous campaign	1 - 871
Previous	Numeric	number of contacts performed before this campaign and for this client	0-275
Poutcome	Categorical	the outcome of the previous marketing campaign	'unknown', 'failure', 'other', 'success'
Y	Numerical	Whether the client subscribed to a term deposit or not	0,1

3.2.Preprocessing

We have checked the dataset to make sure there are no blank or incomplete values. We also examined the dataset for each feature's value consistency. We discovered that a tiny fraction of "unknown" values may be found in employment, education, and interaction. Since one of the stages to prevent over fitting is to verify the Pearson Correlation Coefficient [17] between dataset characteristics before choosing the variables for the regression model, we did so. The range of correlation values is 0 to 1. The connection is thought to be weak between 0 and 0.3, moderate between 0.3 and 0.7, and strong between 0.7 and 1. There is no connection between the independent variables in the figure 1 correlation matrix, which is a strong indication that the dataset's characteristics are not redundant.

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

Figure 1: Correlation Matrix

3.3.Data Exploration and Visualization

The presentation of data exploration and visualization follows. Features in a dataset might be categorical or numerical. Number features include "age," "balance," "length," "campaign," "pdays," and "prior," whereas categorical characteristics include "job," "marriage," "education," "default," housing," "loan," "poutcome," and "y." For categorical features, bivariate analysis and plots of each feature against the goal Y are performed. Bivariate analysis is beneficial since it demonstrates to the contact center that they must focus on a certain consumer group. The correlation between the job characteristic and y is seen in Figure 2. More persons with professional profiles were approached by the bank. Most term depositors are highly qualified compared to other people.

The relationship between the marital characteristic and y is seen in Figure 3. The bank was more interested in married and single persons than divorced ones. The three factors are shown in decreasing order. Samples' direct relationship to the target column. The samples that were more "married" had more subscribers. The relationship between the education characteristic and y is seen in Figure 4. There were more subscribers with advanced degrees. proportional connection. More secondary profiles indicate increased term deposit sales.The relationship between the default feature and y is seen in Figure 5. The overall number of term deposit takers correlates with a high percentage of non-defaulters. People with credit don't appear to want to sign up for a new bank offer, which makes sense.

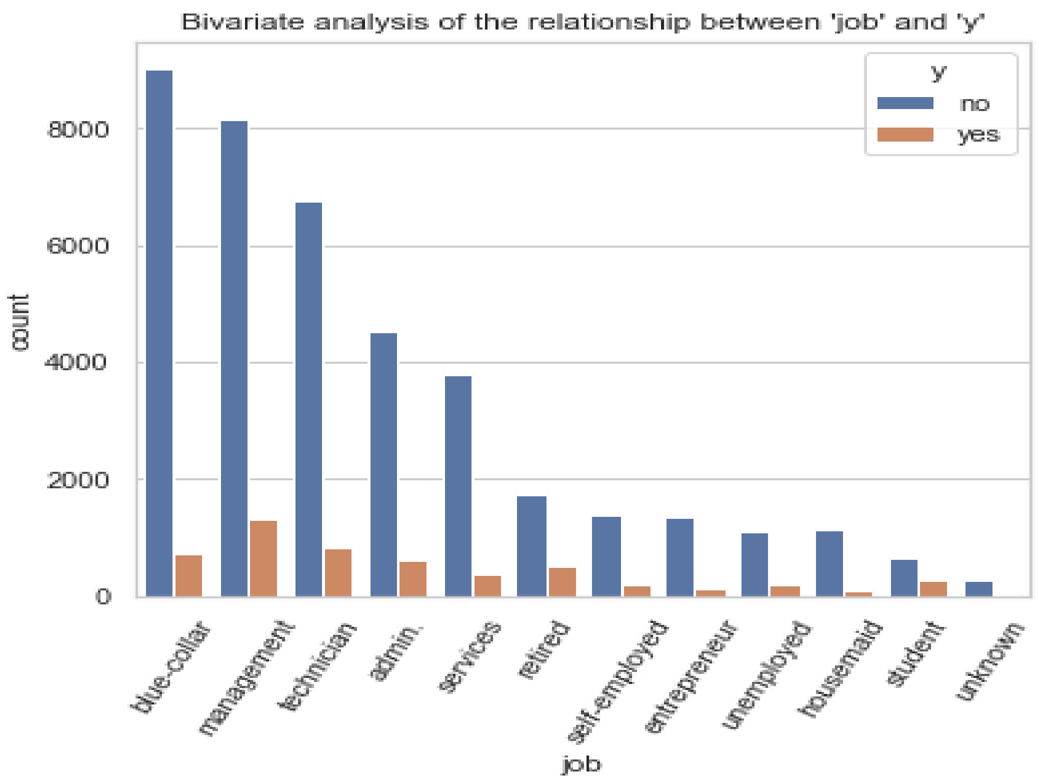

Figure 2: Relationship between job and y.

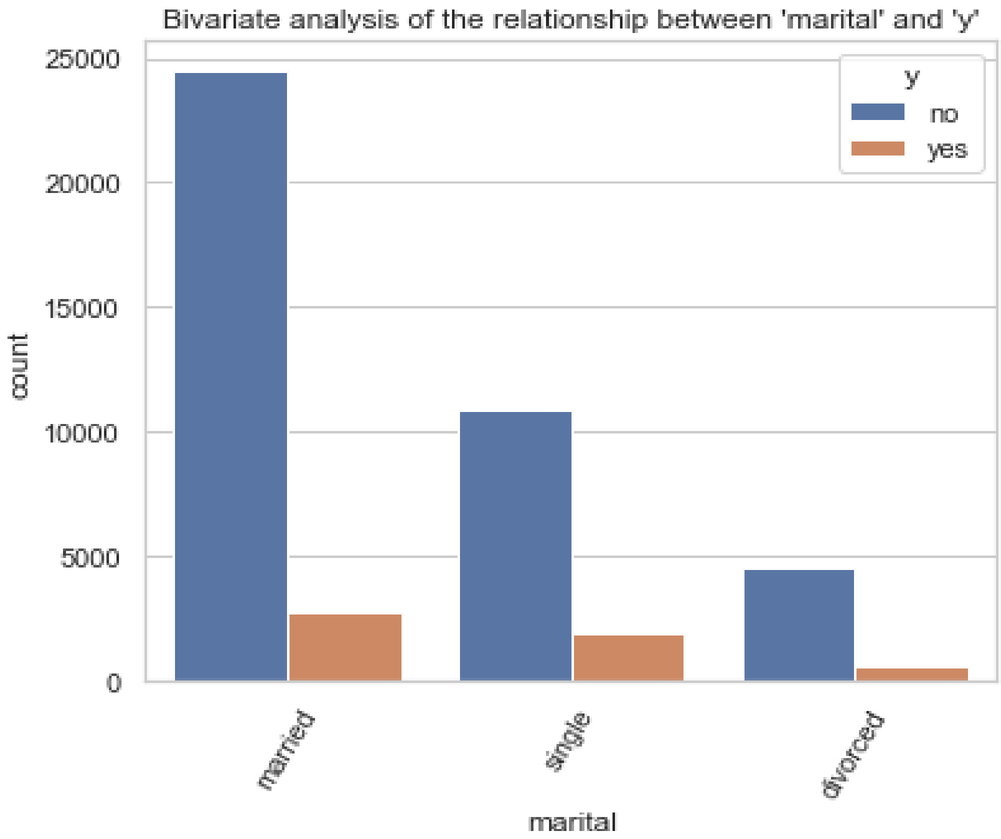

Figure 3: Relationship between marital and y.

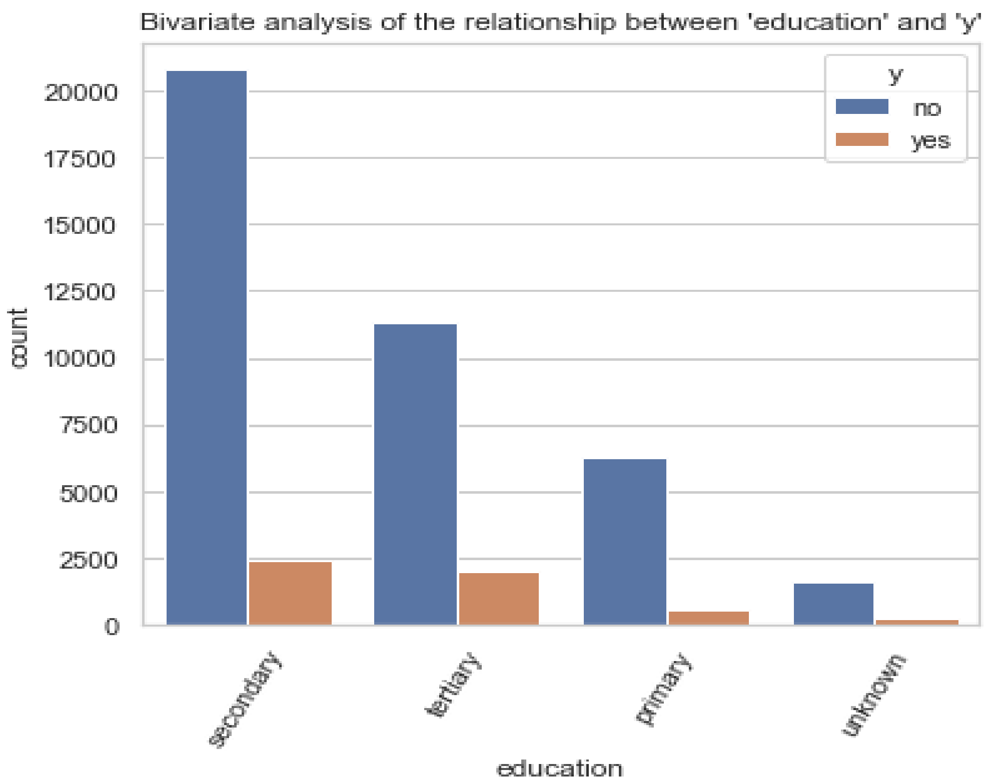

Figure 4: Relationship between education and y.

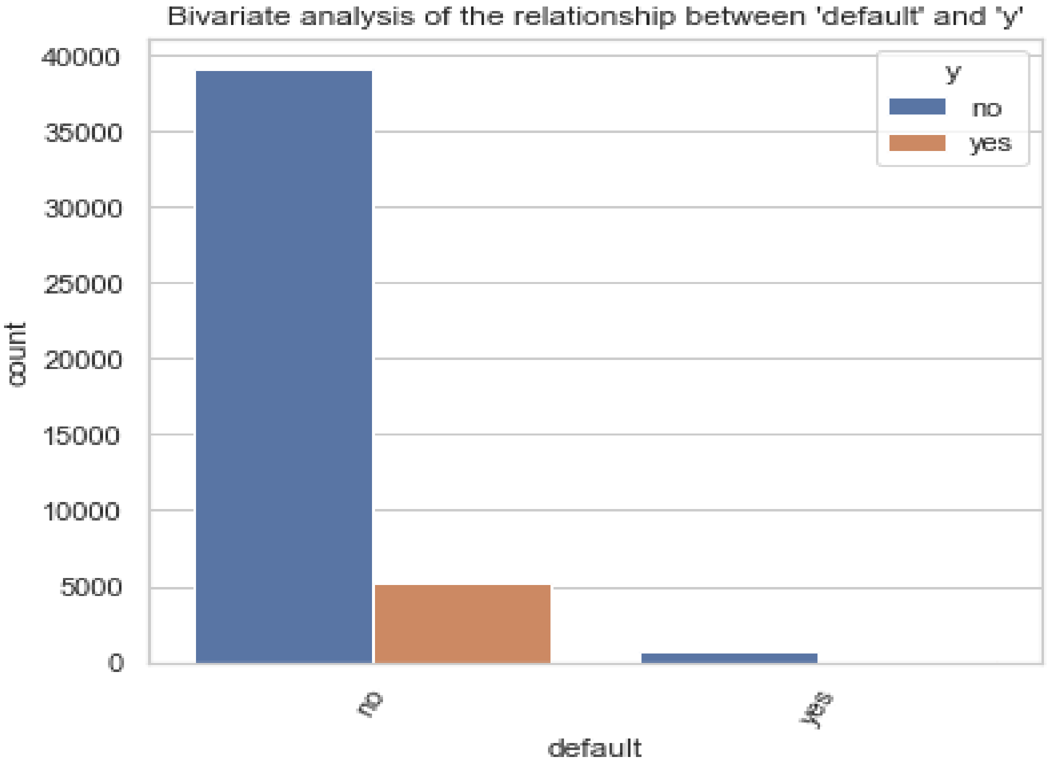

Figure 5: Relationship between default and y.

The relationship between the dwelling descriptive feature and y is seen in Figure 6. More persons have a mortgage. A bigger percentage of those who do not have a mortgage have chosen to enroll in a term deposit. The relative inverse of the percentage.

The relationship between the loan characteristic and y is seen in Figure 7. Similar to house loans, those without personal loans were eager to get a term for a down payment (A higher proportion than housing loans). Only a few borrowers of personal loans decided to subscribe. an exact proportional relationship.

The relationship between the contact feature and y is seen in Figure 8. In this graph, the direct ratio demonstrates that more customers who were contacted through cellular were enrolled in a deposit term. The "unknown" variable will be processed together with the other variables using an imputation approach.

The relationship between the month feature and y is seen in Figure 9. Comparing May to the previous months, there were a few more subscriptions. Except for December and January, the subscription average is almost constant regardless of how many individuals are contacted. There were the fewest subscribers throughout these months. One cause may be that people travel for vacations. (In the Americas, people are used to taking vacations around this time of year.) Given that the plot depicts a proportionate distribution of "yes," the "month" feature will be eliminated since it does not affect the result.

Vol.10 Issue 7, Pg.: 1-20
July 2022

The relationship between the p outcome feature and y is seen in Figure 10. This one is in line with the results of the earlier marketing campaign's success. What does "unknown" mean? It indicates that 78.7% of those contacted were unaware of the preceding marketing initiative. The success of this campaign may benefit the previous campaign. Although "success" has a relatively low proportion, it is nevertheless important for the study.

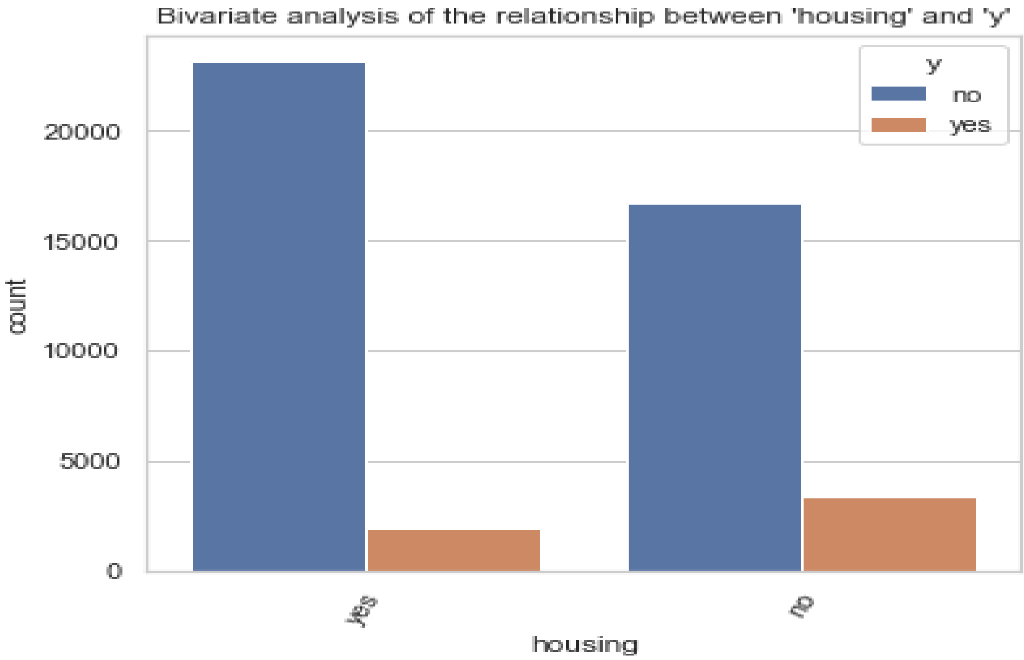

Figure 6: Relationship between education and y

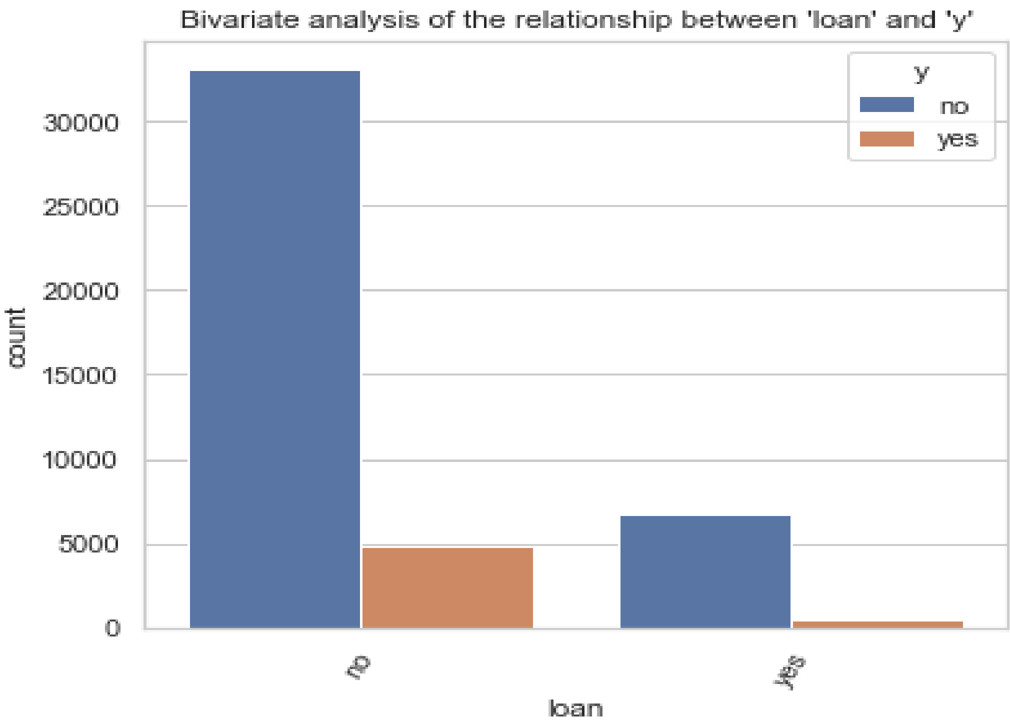

Figure 7: Relationship between loan feature and y

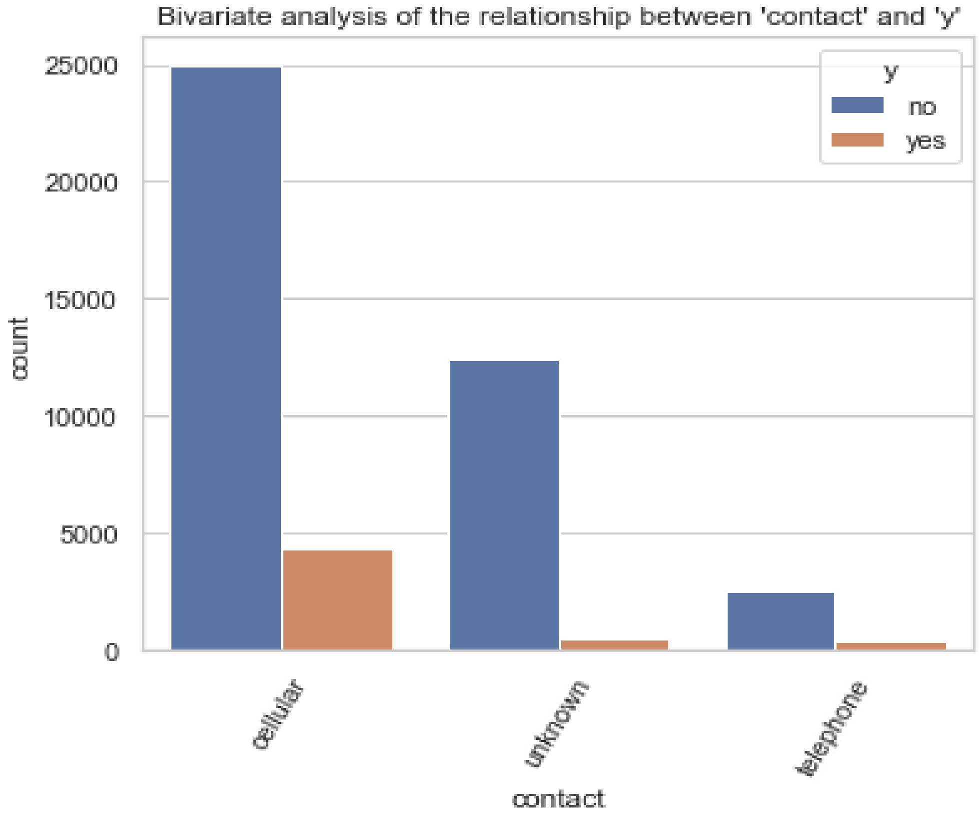

Figure 8: Relationship between contact and y.

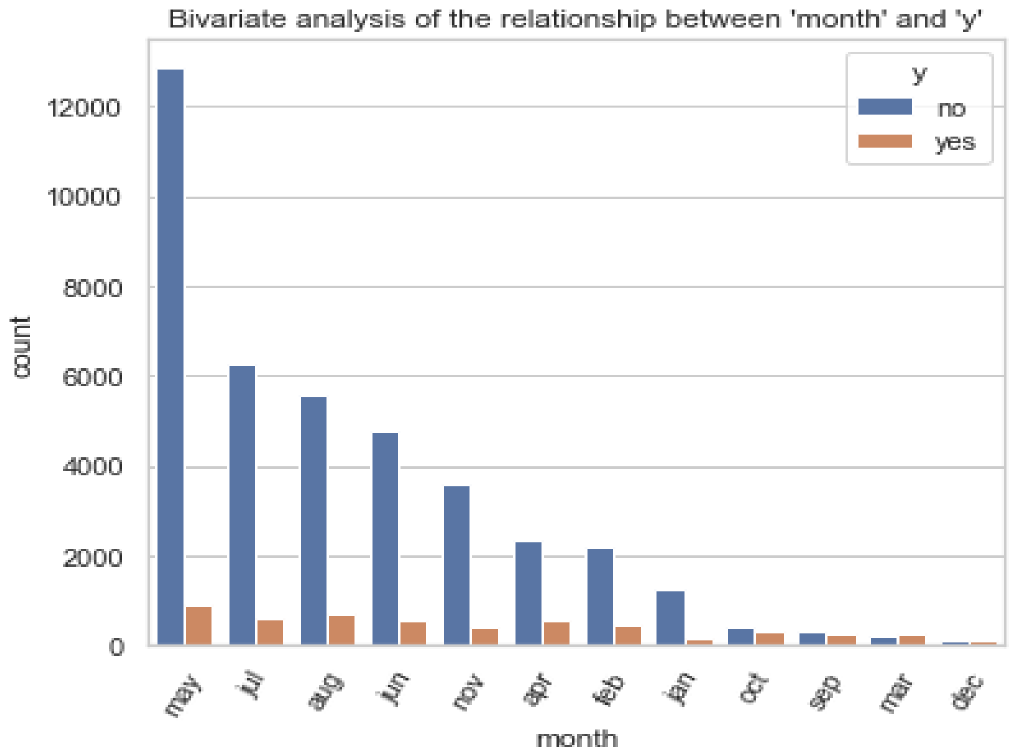

Figure 9: Relationship between month and y

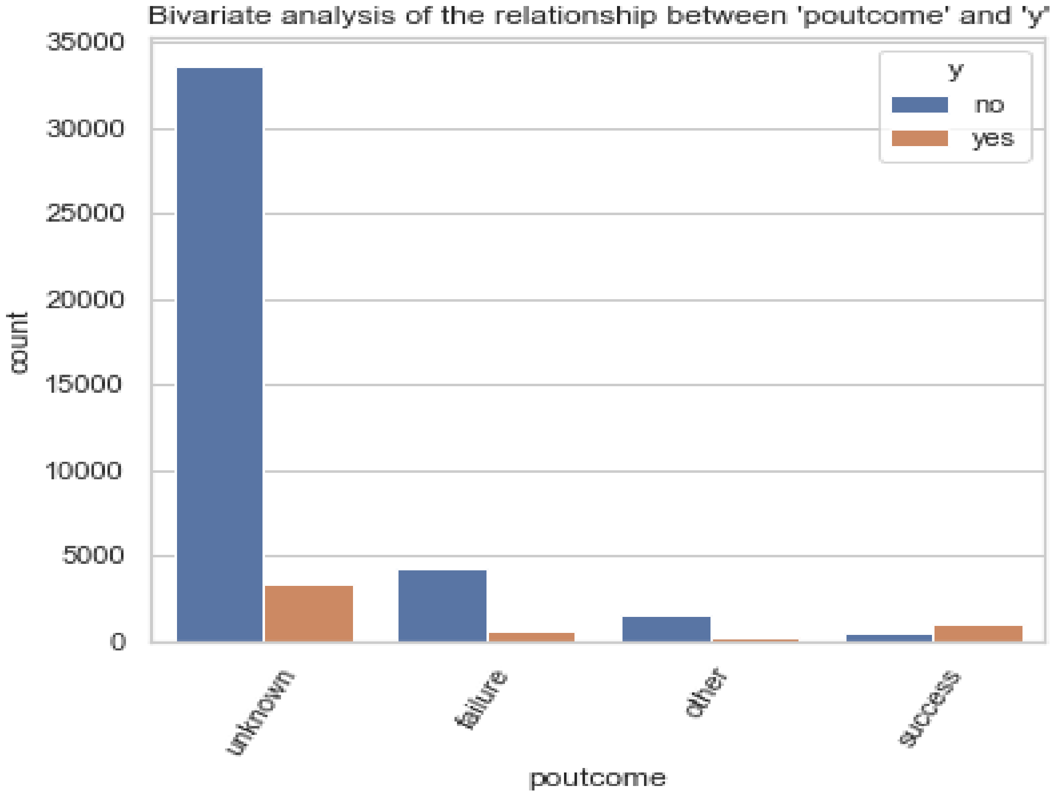

Figure 10: Relationship between p outcome and y

Univariate analysis for the target y is shown in Figure 11. Many of the persons who were contacted expressed a desire not to sign up for the bank's offering. Only 13,2 percent of the population agreed. This significant disparity reveals an issue with class inequality.

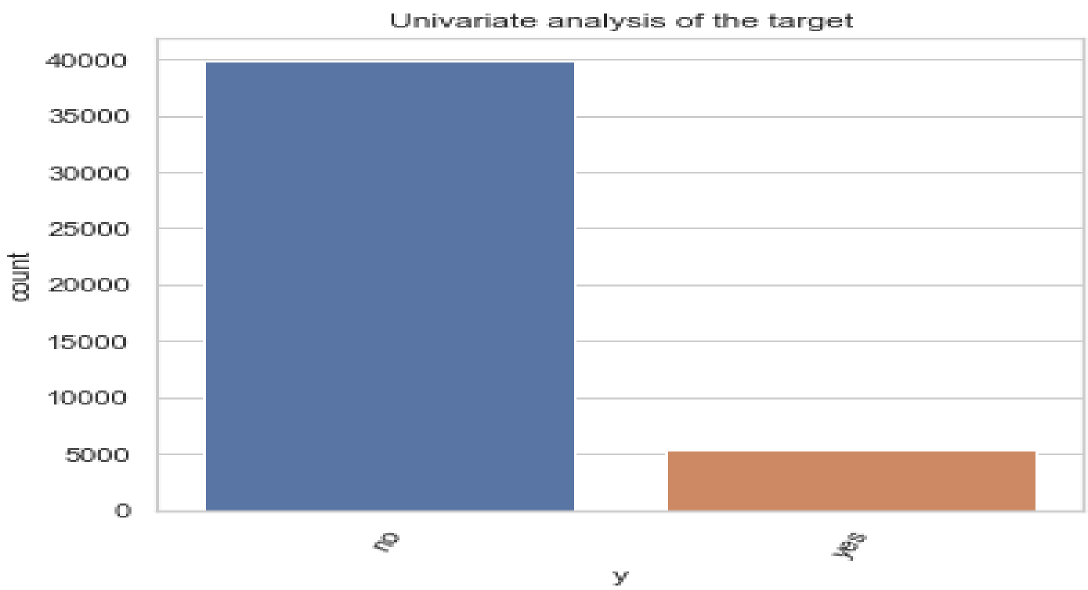

Figure 11: Univariate analysis for target y

Vol.10 Issue 7, Pg.: 1-20
July 2022

Box-plot analysis and histograms [18] are shown for the numerical characteristics. Additionally, the median, mean, and mode [19] for each characteristic are displayed. An analysis of the age feature is shown in Figure 12. This function uses box visual representations to illustrate the connection between "age" and the categorized target variable. Additionally, the histogram shows a bell-shaped picture with a left-shifted normal distribution (shown below). The age range of the population is 20 to 60. The anticipated age range of the box plot is between 30 and 50. Probably because people are more steady and productive at this time. If we examine the job feature, this tendency becomes stronger.

An analysis of the balancing characteristic is shown in Figure 13. A median of 0 is shown on the green line. This indicates that the majority of those contacted had annual balances that are almost $0. An analysis of the day characteristic is shown in Figure 14. The histogram displays some degree of symmetry with a peak on day 20 over the whole data set. This characteristic won't be included in the dataset analysis since it doesn't significantly affect the results. People may subscribe any day of the week, according to plots.

An analysis of the duration characteristic is shown in Figure 15. The "y" result is impacted by this feature. Below, we can observe that the majority of consumers reject the offer when the time is between 0 and the first two minutes. The decision for the residual samples ranges from more than 2 minutes to 12 minutes. Few people spend a long time accepting or rejecting the offer.

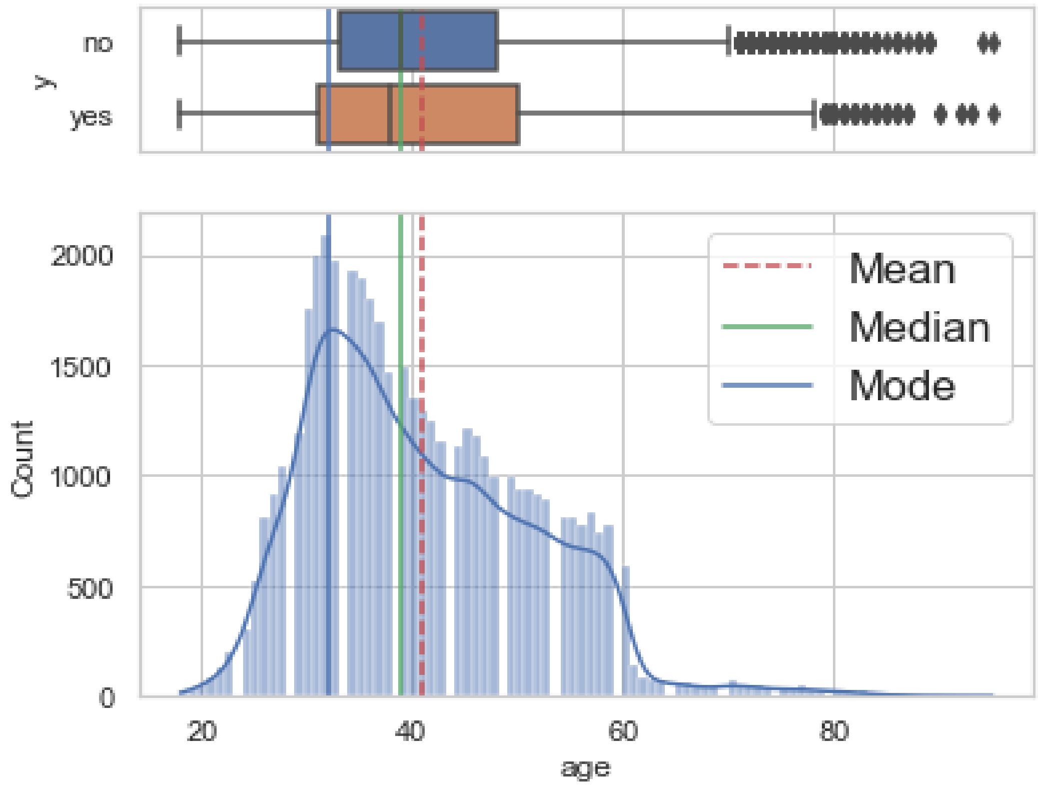

Figure 12: Age feature analysis

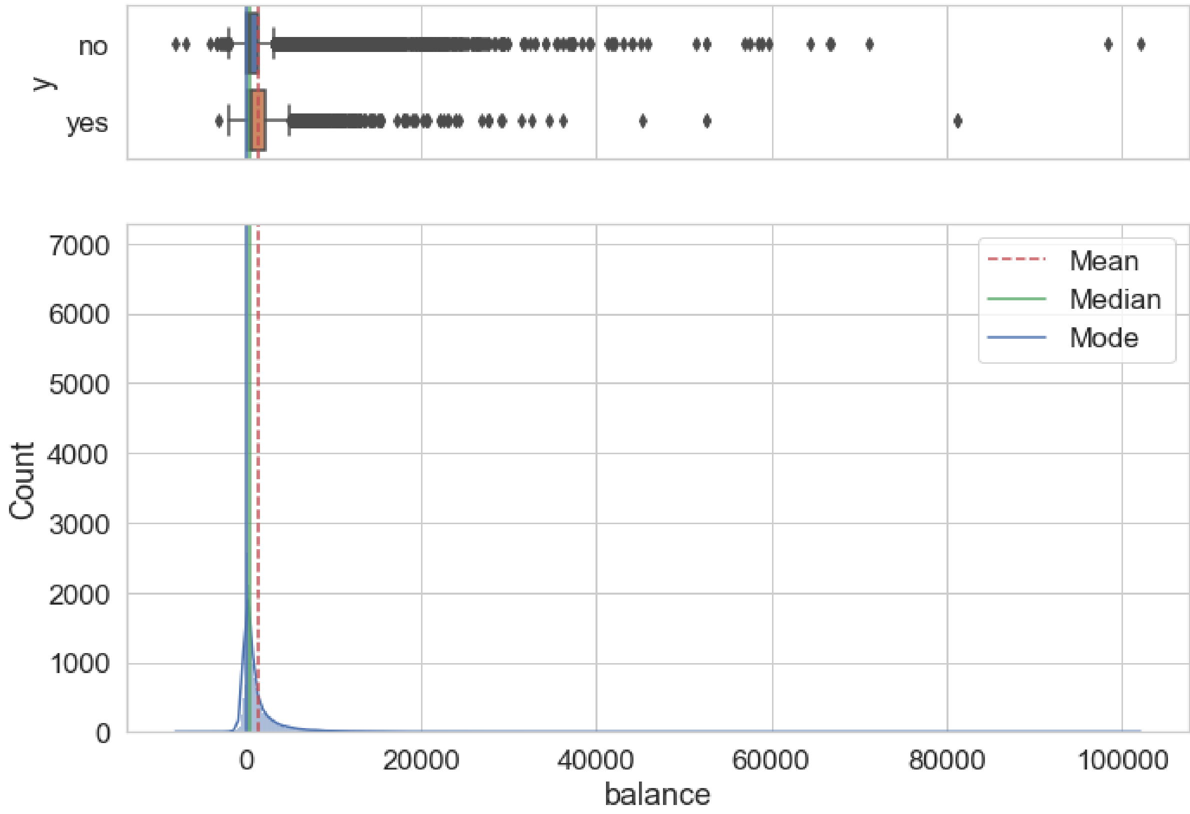

Figure 13: Balance feature Analysis.

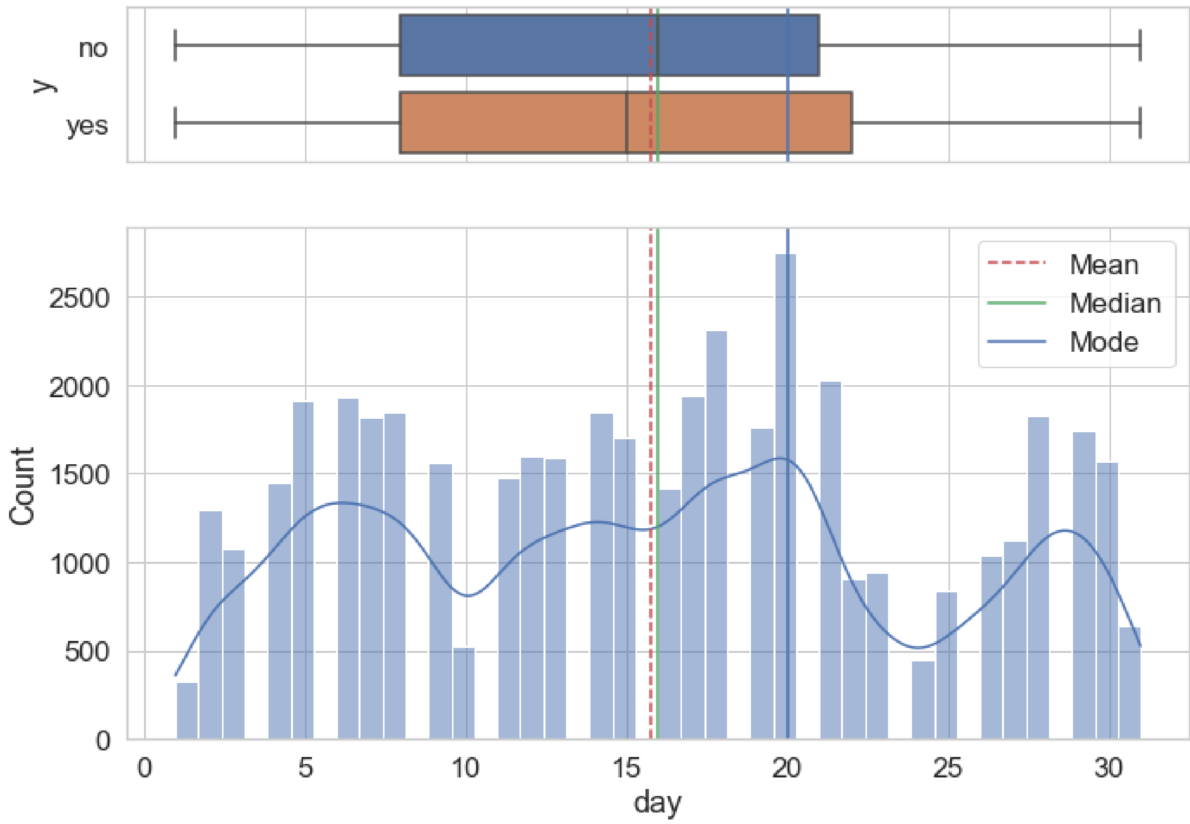

Figure 14: Day feature analysis.

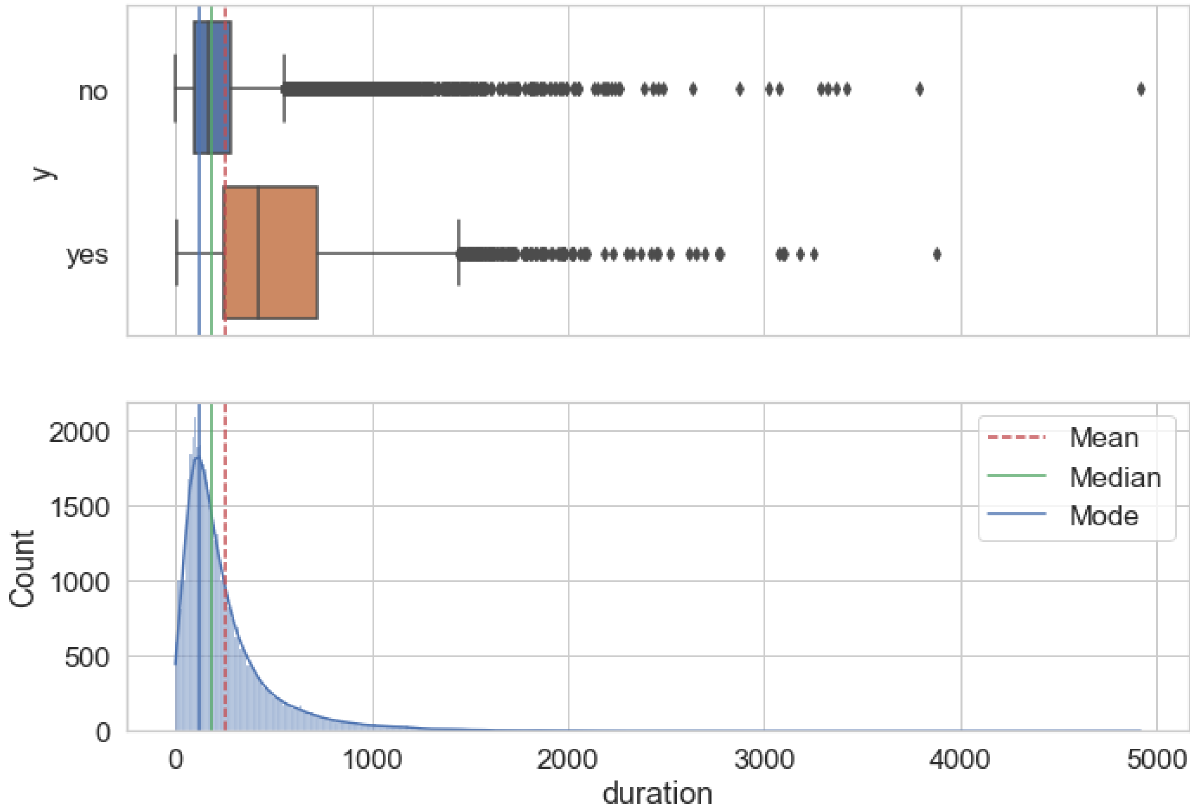

Figure 15: Duration feature analysis.

An analysis of the campaign characteristic is shown in Figure 16. The term "campaign" refers to the total number of contacts made throughout this campaign. The plots below show that the "yes" and "no" are evenly split. Second, those who were often approached were the ones who chose to sign up for a deposit period. In contrast, making repeated attempts to get in touch with individuals is a waste of time.

An analysis of the pdays characteristic is shown in Figure 17. The amount of days since the customer was last reached by a prior campaign makes up for this characteristic (numeric, -1 means the client was not previously contacted). The box plot shows that almost all of the participants were contacted for the first time, and the median is -1. As a result, this feature will be eliminated since it has no impact on the result.

An analysis of the preceding characteristic is shown in Figure 18. "Prior" refers to the total amount of contacts made for this customer and before this campaign. As stated below, 36954 are associated with 0. It indicates that 36954 people were originally contacted for this campaign. Additionally, there is no box-plot, which means there is no distribution and no link to the aim. It will also be excluded from the analysis.

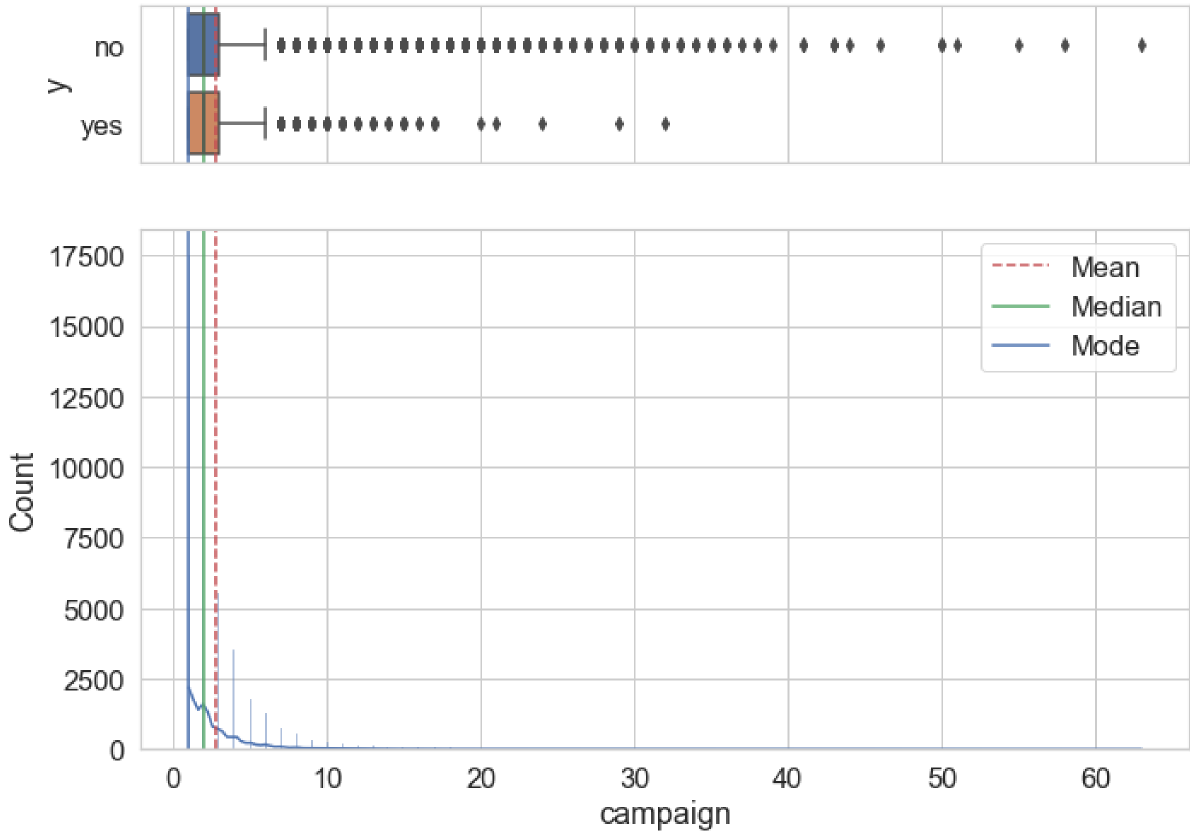

Figure 16: Campaign feature analysis.

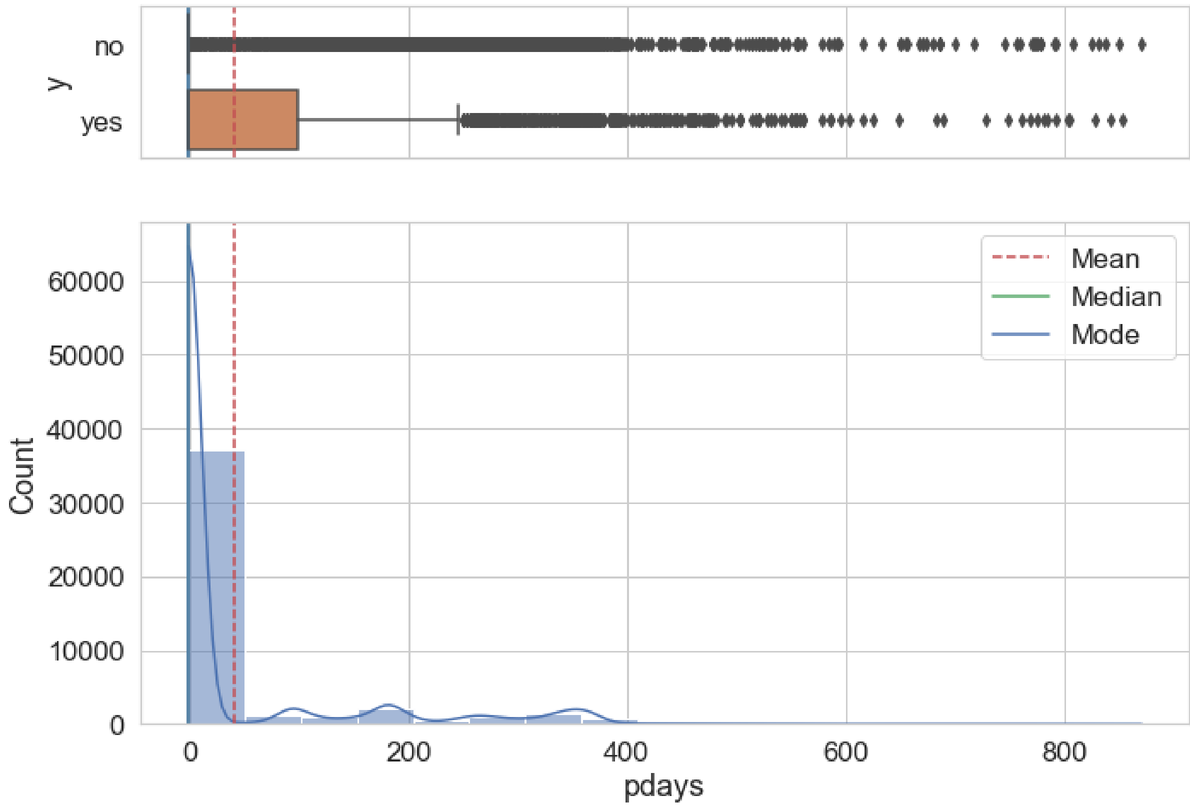

Figure 17: pdays feature analysis.

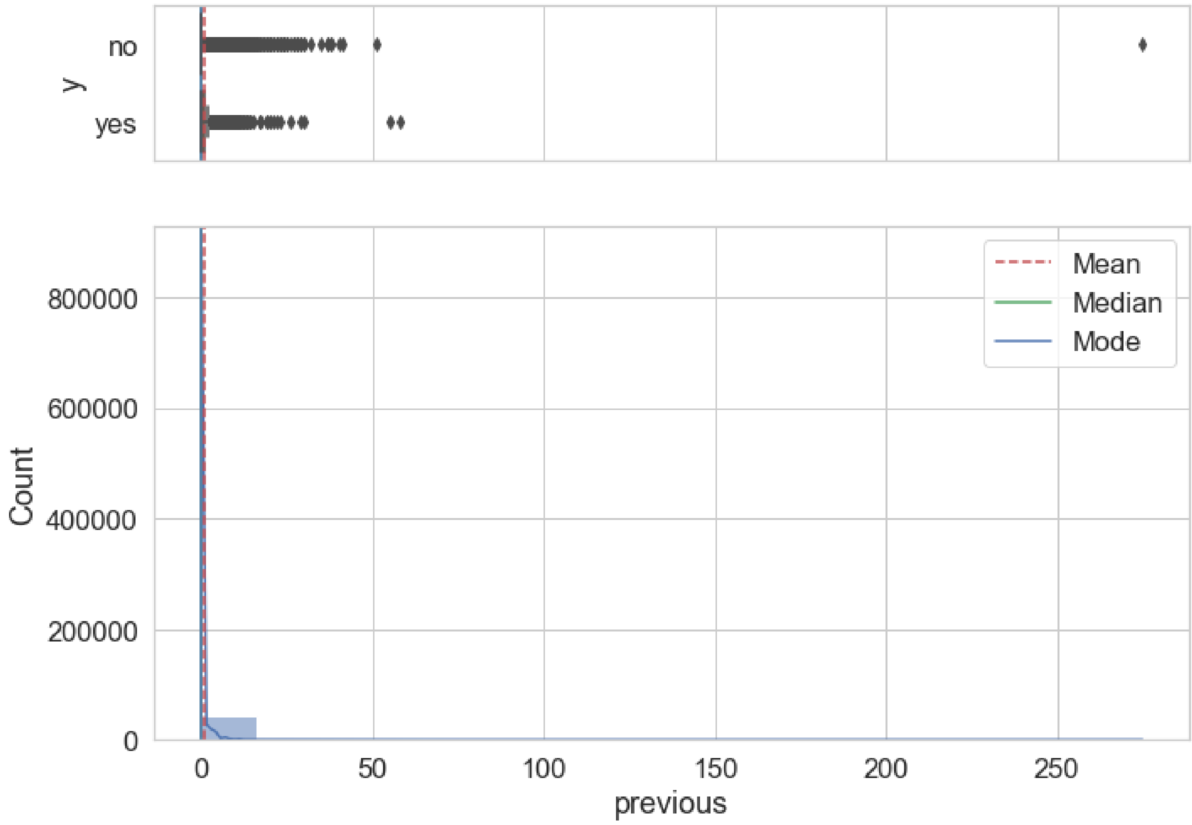

Figure 18: previous feature analysis.

3.4.Prediction

We must convert category variables into numerical values since ML models need numerical input. We used One hot encoding [20] for the following categorical features 'job', 'marital', 'default', 'housing', 'loan', 'contact', 'month', and 'poutcome'. For the “education” feature, we used label encoding [20] where we labeled 'unknown' to 0, 'primary' to 1,'secondary' to 2, and 'tertiary' to 3. Also, the target variable y is encoded as 'no' to 0 and 'yes'to 1. The dataset is originally split into training and testing but the dataset is unbalanced so we needed to consider that while training so we used the balanced class weight approach [21] and 10-fold cross-validation [22] to solve the imbalanced class problem. A random forest [23] algorithm is used for training and testing and a perfect classifier is achieved with 99.5% accuracy. Figure 19 shows a comparison between the proposed system and other state-of-the-art systems that used bank marketing datasets. The system outperformed all the state-of-the-art techniques and achieved perfect classification.

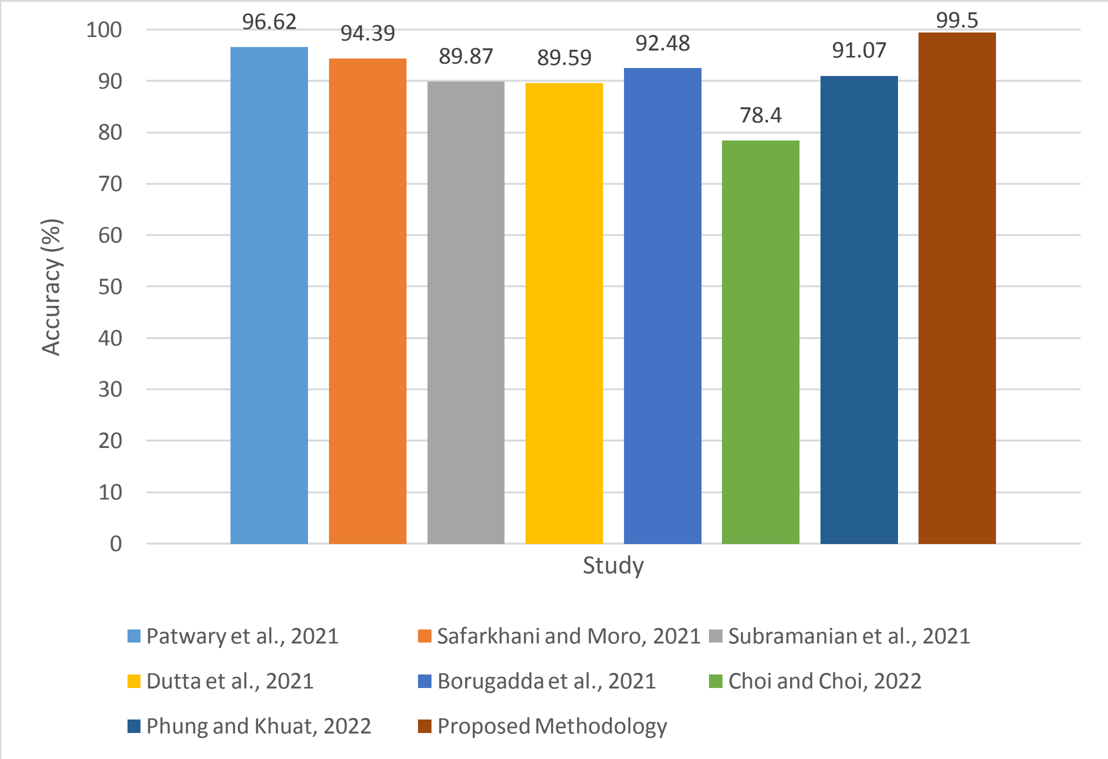

Figure 19: Comparison between systems using bank marketing dataset.

4. Conclusion

One of the most important concerns for every financial firm is bank deposits. Analyzing relevant data to forecast if a consumer will be a depositor is quite difficult. Recent studies have shown that corporate organizations and the banking industry are badly impacted by the economic downturn and the economy's ongoing decline. Our research reveals the remarkable influence that machine learning methods may have on the outcome of a telemarketing campaign. Data preparation and model assessment are the two main phases. In the first phase, data must be cleaned by removing duplicate records and determining if missing values should be kept or removed, data visualization and response encoding approach to encode category characteristics using label and one-hot encoding. The dataset was initially divided into training and testing, but since the dataset was uneven, we wanted to take it into account when training. To address the imbalanced class issue, we employed the balanced class weight strategy and 10-fold cross-validation. A perfect classifier is created using the Random Forest method for both training and testing. The proposed method obtained perfect classification and outperformed all state-of-the-art methods.

References

[1] Han, J., Kamber, M., & Pei, J. (2011). Data Transformation by Normalization. Data Mining: Concepts and Techniques.

[2] Simon, J. P. (2019). Artificial intelligence: scope, players, markets and geography. *Digital Policy, Regulation and Governance*, *21*(3), 208-237.

[3] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Vol.10 Issue 7, Pg.: 1-20
July 2022

[4] Kose, U. (2019). Using artificial intelligence techniques for economic time series prediction. In *Contemporary Issues in Behavioral Finance*. Emerald Publishing Limited.

[5] Patwary, M. J., Akter, S., Alam, M. B., & Karim, A. R. (2021). Bank Deposit Prediction Using Ensemble Learning. *Artificial Intelligence Evolution*, 42-51.

[6] Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22-31.

[7] Safarkhani, F., & Moro, S. (2021). Improving the Accuracy of Predicting Bank Depositor's Behavior Using a Decision Tree. *Applied Sciences*, *11*(19), 9016.

[8] Subramanian, R. S., Prabha, D., Aswini, J., Maheswari, B., & Anita, M. (2021, February). Alleviating NB conditional independence using multi-stage variable selection (MSVS): banking customer dataset application. In *Journal of Physics: Conference Series* (Vol. 1767, No. 1, p. 012002). IOP Publishing.

[9] Dutta, S., Bose, P., Goyal, V., Bandyopadhyay, P., & Kumar, S. (2021). Applying convolutional-GRU for term deposit likelihood prediction. *International Journal of Engineering and Management Research*, *11*.

[10] Borugadda, P., Nandru, P., & Madhavaiah, C. (2021). Predicting the success of bank telemarketing for selling long-term deposits: An application of machine learning algorithms. *St. Theresa Journal of Humanities and Social Sciences*, *7*(1), 91-108.

[11] Choi, Y., & Choi, J. (2022). How does Machine Learning Predict the Success of Bank Telemarketing?.

[12] Phung, T. D., & Khuat, D. B. (2022). *Potential Customers Prediction in Bank Telemarketing* (Doctoral dissertation, FPTU Ha Noi).

[13] Hodgson, G. M. (2011). The great crash of 2008 and the reform of economics. In *The Handbook of Globalisation, Second Edition*. Edward Elgar Publishing.

[14] Keeley, M. C. (1990). Deposit insurance, risk, and market power in banking. *The American economic review*, 1183-1200.

[15] Zhuang, Q. R., Yao, Y. W., & Liu, O. (2018). Application of data mining in term deposit marketing. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 2).

[16] Palaniappan, S., Mustapha, A., Foozy, C. F. M., & Atan, R. (2017). Customer profiling using classification approach for bank telemarketing. *JOIV: International Journal on Informatics Visualization*, *1*(4-2), 214-217.

[17] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1-4). Springer, Berlin, Heidelberg.

[18] Barnett, O., & Cohen, A. (2000). The histogram and boxplot for the display of lifetime data. *Journal of Computational and Graphical Statistics*, *9*(4), 759-778.

[19] Kaas, R.,& Buhrman, J. M. (1980). Mean, median and mode in binomial distributions. *Statistica Neerlandica*, *34*(1), 13-18.

[20] Rodríguez, P., Bautista, M. A., Gonzalez, J., & Escalera, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing*, *75*, 21-31.

[21] King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, *9*(2), 137-163.

[22] Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*(2), 137-146.

[23] Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, *47*(1), 31-39.