International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Customer Deposit Forecasting - Optimizing Deposit Predictions through CRM and Machine Learning Integration

Sreeja S
B.Tech
School of Engineering -AIML
Mallareddy University, India
2111cs020541@
mallareddyuniversity.ac.in

Sreeja B
B.Tech
School of Engineering - AIML
Mallareddy university, India
2111cs020542@
mallareddyuniversity.ac.in

Sreeja B
B.Tech
School of Engineering - AIML
Mallareddy university, India
2111cs020543@
mallareddyuniversity.ac.in

Sreeja R
B.Tech
School of Engineering - AIML
Mallareddy university, India
2111cs020544@
mallareddyuniversity.ac.in

Sreeja U
B.Tech
School of Engineering - AIML
Mallareddy university, India
2111cs020545@
mallareddyuniversity.ac.in

Sreenidhi M
B.Tech
School of Engineering - AIML
Mallareddy university, India
2111cs020546@
mallareddyuniversity.ac.in

K.Manoj Sagar M.E
Assistant Professor / AI & ML
School of Engineering
Malla Reddy University, Telangana
manojsagar_k@mallareddyuniversity.ac.in

***Abstract*:** This machine learning project revolves around predicting customer deposits within a banking context. The banking sector heavily relies on revenue generated from long-term deposits by customers. Understanding customer characteristics is crucial for banks to enhance product sales, leading to the employment of marketing strategies aimed at target customers. The advent of data-driven decisions has prompted the utilization of data analysis, feature selection, and machine learning techniques to analyze customer characteristics and predict their decisions accurately. Leveraging a comprehensive dataset encompassing customer demographics, financial histories, and interaction records, the project employs thorough exploratory data analysis (EDA) and preprocessing techniques to unveil patterns and relationships. Categorical features like job, marital status, and education are scrutinized, providing valuable insights. Preprocessing steps include encoding categorical variables, addressing outliers, and handling imbalances in the dataset. The dataset is partitioned into training and testing sets, and machine learning models such as Random Forest Classifier and XG Boost Classifier are employed for prediction, optimized through GridSearchCV. Cross-validation scores reveal the models' efficacy, with Random Forest and XGBoost demonstrating promising performance. This project underscores the significance of meticulous data analysis and preprocessing in constructing accurate predictive models within the banking domain. The identified strengths of our model suggest its potential application in optimizing customer deposit forecasts. This contribution advances the field of financial predictive modeling by introducing a robust solution tailored for accurate customer deposit prediction, thereby facilitating informed decision-making in the banking industry.

Keywords:-Random Forest Classifier, XGB Classifier, GridSearch CV, XGBoost

I.INTRODUCTION

In the realm of bank marketing, the strategic deployment of campaigns plays a pivotal role in influencing consumer behaviour and driving desired outcomes[1]. Conventionally, two approaches have dominated this landscape: mass marketing targeting the broader population and targeted marketing focusing on specific demographic segments. A notable challenge arises in the efficacy of mass marketing, as studies reveal a comparatively shallow positive response rate for product purchase or service subscription when compared to targeted marketing efforts[1]. The inherent inefficiencies in mass campaigns, despite their contribution to sales, underscore the need for a more precise and resource-efficient strategy. Direct marketing, exemplified by techniques such as indirect telemarketing, proves effective in selling products by directly engaging potential customers through phone calls. However, the difficulty lies in identifying and reaching the right audience within a specific group.[4]

The advent of data-driven decision-making has ushered in a transformative era for marketing managers. Statistical strategies now empower them to discern potential buyers with greater accuracy, minimizing the challenges associated with identifying customers likely to invest in the bank .This paradigm shift becomes particularly crucial in the current economic landscape, where banks, grappling with economic turmoil, seek to bolster their financial reserves through the sale of long-term deposits[1]. The confluence of economic uncertainties and the imperative to sell long-term stakes necessitates a strategic overhaul for marketing managers[4]. The focus is not merely on increasing sales but on optimizing

the positive response rate while efficiently managing scarce resources. In pursuit of this objective, bank marketing managers turn to sophisticated multivariate data classification methods, leveraging insights gained from previous campaigns to identify and target potential customers effectively.

This paper delves into the evolving dynamics of bank marketing, accentuating the need for a paradigm shift in campaign strategies. By exploring the intersection of statistical methodologies and marketing objectives, we aim to provide insights into the potential of multivariate data classification methods in identifying prospective customers. The subsequent sections detail the methodology employed, the contextual backdrop of economic challenges, and the implications of adopting a data-driven approach in the domain of bank marketing.

II.LITERATURE REVIEW

The landscape of bank marketing has witnessed a dichotomy in campaign strategies, primarily characterized by mass and targeted approaches[1]. Mass marketing, often directed at the general population, has historically contributed to sales but suffers from a notable drawback—the positive response rates are comparatively shallow, leading to inefficiencies in resource allocation .In contrast, targeted marketing, focusing on specific demographic segments, has shown promise in yielding higher positive response rates. Direct marketing techniques, such as indirect telemarketing, represent a more precise approach by directly engaging potential customers through personalized communication channels[4]. However, the challenge lies in identifying the right audience within specific demographic groups. The integration of data-driven decision-making into the marketing landscape has emerged as a transformative force[1]. Marketing managers are increasingly turning to statistical strategies to overcome the limitations of traditional approaches. These data-driven methodologies empower managers to discern potential buyers more accurately, a particularly valuable asset in the face of economic uncertainties. The contemporary economic landscape, marked by turmoil in various countries, has placed additional demands on banks to sell long-term deposits and bolster their financial reserves . This scenario intensifies the pressure on marketing managers to devise strategies that not only increase sales but also optimize positive response rates. Multivariate data classification methods present a promising avenue for achieving this delicate balance[6].

However, a critical examination of existing literature reveals a research gap in comprehensively exploring the intersection of statistical methodologies and marketing objectives in the context of bank marketing[1]. While some studies highlight the efficacy of targeted approaches, there is a paucity of research that systematically investigates the application of multivariate data classification methods in optimizing positive response rates[7].This paper aims to bridge this gap by providing a nuanced understanding of the evolving dynamics in bank marketing. Through a synthesis of existing literature, we identify the need for a paradigm shift in campaign strategies and propose an exploration of multivariate data classification methods as a solution[9]. The subsequent sections of this paper delve into the methodology, economic backdrop, and implications of adopting a data-driven approach in the domain of bank marketing[1].

III. PROBLEM STATEMENT

Bank marketing faces challenges in campaign optimization, marked by limited positive response rates in mass marketing and imprecise audience identification in targeted efforts. Incomplete integration of data-driven strategies and a gap in exploring multivariate data methods further compound the issues. Economic challenges, especially the need to sell long-term deposits, are often overlooked.

To address these, our research introduces a precision-focused machine learning model, seamlessly integrating data-driven strategies, offering a robust framework for marketing managers to leverage insights effectively. The project utilizes a carefully curated dataset from previous marketing campaigns, incorporating diverse variables to train and evaluate the model. This dataset forms the basis for addressing the research questions and hypotheses guiding the project. The central research questions revolve around the efficacy of the machine learning model in improving positive response rates, its comparative performance against existing methods, and its adaptability to varying economic contexts.

Hypotheses center around assessing the efficacy of the proposed machine learning model in optimizing positive response rates compared to traditional methods, evaluating its comparative performance, and exploring its adaptability to varying economic contexts. Additionally, hypotheses focus on the model's ability to enhance precision in audience identification through multivariate data methods and its impact on long-term deposit sales.

IV. METHODOLOGY

The primary objective of this research is to develop a robust Machine Learning model for categorizing clients into two distinct groups: those inclined to subscribe to long-term deposits and those who are not. Addressing the challenges of a high-dimensional and imbalanced dataset, traditional sampling techniques and algorithm-based approaches may prove insufficient. To overcome class imbalance, we opt for a meticulous feature selection process, emphasizing Exploratory Data Analysis (EDA) rather than Principal Component Analysis (PCA) to enhance interpretability without losing crucial information. Leveraging factor analysis to discern associations, we employ Logistic Regression, Random Forest, Support Vector Machine (SVM), Decision Tree, and XG Boost classifiers for predicting term deposit subscription. Additionally, the study endeavors to construct a reliable recommendation algorithm based on client characteristics. The binary target variable, indicating subscription preference, guides our use of multiple classifiers, and their accuracy is rigorously compared to determine the most effective model. Correlation analysis is instrumental in unveiling rules for banks to identify potential subscribers, considering factors such as age, marital status, gender, and education. Our exploration extends to evaluating the impact of personalized bank campaign efforts, including the number of phone calls and previous campaign results, to unveil patterns enhancing marketing strategies and services, thus

contributing to the dynamic landscape of customer relations in the banking sector.

A. Dataset Overview

This dataset, sourced from the UCI Machine Learning Repository, sheds light on a financial institution's marketing campaign[1]. It holds a trove of information vital for dissecting and refining future marketing strategies. The data covers diverse aspects, including age, job type, marital status, education background, and financial details. Categorical features span job roles, marital and education statuses, credit default, housing and personal loans, contact methods, and temporal details like the last contact day and month. Numeric attributes capture individual balances, last contact duration, campaign interaction counts, and outcomes of previous marketing efforts. The target label, 'deposit,' denotes a binary outcome ('yes' or 'no') indicating whether a client has subscribed to a term deposit. This dataset lays the foundation for in-depth analysis, guiding the formulation of predictive models to decipher the factors influencing term deposit subscriptions.

B. Data Pre-processing

To increase data quality, the primary dataset was pre-processed. Preprocessing techniques were applied to several attributes. The data preprocessing phase begins with an examination of the dataset for unwanted columns, missing values, and features with only one value—no such issues are identified. Focusing on categorical data, an exploration reveals nine categorical features, with 'job' and 'month' having the highest categorical values. Further analysis showcases noteworthy patterns, such as a prevalence of clients in management roles, a higher representation of married individuals, and a concentration of secondary education backgrounds. The 'default' feature exhibits a skewed distribution, leaning towards 'no,' suggesting its limited impact and potential for removal. Relationships between categorical features and the label ('deposit') unveil insights, such as retired clients showing a high interest in deposits and clients with housing loans displaying lower interest. Success in the previous campaign ('poutcome=success') significantly increases the likelihood of deposit interest. Exploration of numerical features reveals seven variables with no discrete values. Continuous numerical features like age, days, balance, duration, campaign, pdays, and previous interactions exhibit diverse distributions. Notably, clients showing interest in deposits tend to have longer interaction durations. Outlier analysis identifies anomalies in age, balance, duration, campaign, pdays, and previous features. The correlation matrix highlights no strong correlations between numerical features. Importantly, the dataset demonstrates a balanced distribution. These preprocessing insights set the stage for refining and optimizing the dataset to bolster the effectiveness of subsequent machine learning models.

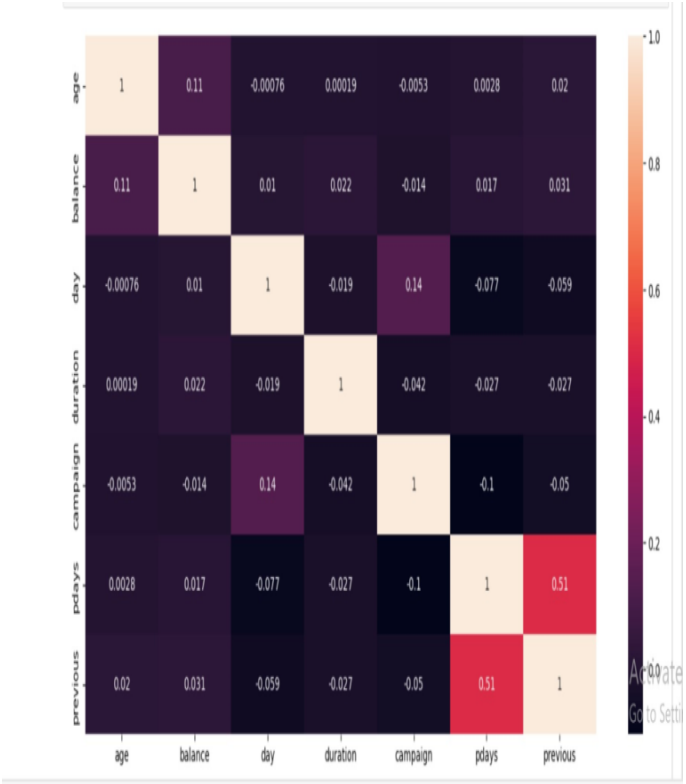

Fig 1: Correlation Analysis

C. Machine Learning Algorithms

Random Forest(RL) Classifier

Random forest is a supervised learning algorithm that can classify and predict data. This ensemble method harnesses the collective decision-making power of multiple trees. Examining features like age, job type, and balance, it consolidates diverse insights to predict whether a client will deposit money. Picture it as consulting a group of friends, each adept at decision-making in different scenarios. Random Forest amalgamates these decisions, offering a more accurate and reliable choice.

XG Boost(XG) Classifier

XG Boost is a highly efficient machine learning algorithm known for its speed and accuracy in predictive modeling. This algorithm, akin to a diligent student, refines its predictive prowess with each iteration. By scrutinizing features like call duration, previous contacts, and balances, XGBoost makes nuanced predictions. Imagine a student learning from test mistakes, honing their understanding of complex relationships between factors. XGBoost excels, especially when faced with intricate patterns in the data.

Logistic Regression(LG) Classifier

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.

Decision Tree Classifier

A Decision Tree is a hierarchical model that sequentially poses a series of questions, akin to a game of 20 Questions, to classify data. Examining features individually, such as job type, marital status, and housing, it navigates through a binary tree structure, making decisions based on the most discriminative features at each node. This method yields interpretable results, breaking down complex decisions into a series of intuitive and transparent steps.

Support Vector Machine(SVM) Classifier

Support Vector Machine (SVM) is a powerful machine learning algorithm for classification and regression tasks. It excels in solving complex problems by identifying the optimal hyperplane that separates different classes. This hyperplane, determined through intricate analysis of features like education, balance, and contact types, provides a clear boundary between data points. SVM is particularly effective in scenarios where data exhibits intricate relationships, offering a sophisticated solution with a focus on maximizing the margin between classes for robust generalization.

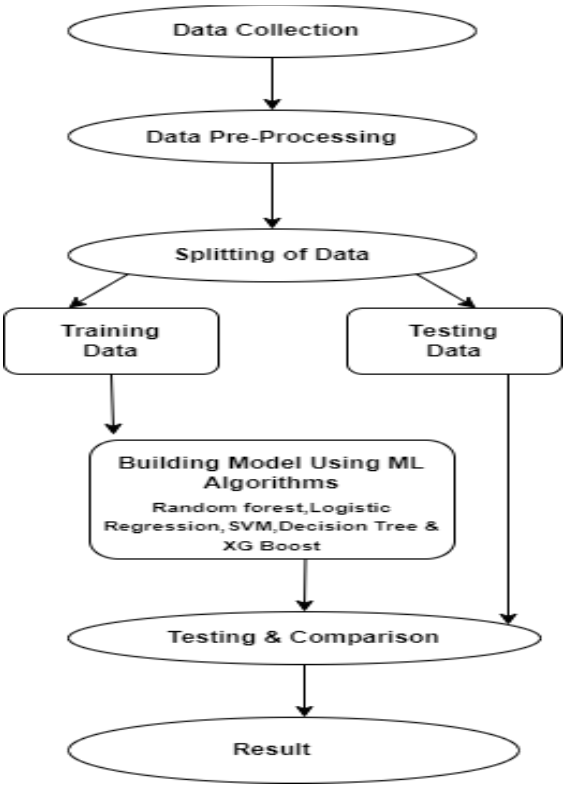

Fig 2: Flow Diagram

V.EXPERIMENTAL RESULT

In our exploration of machine learning tools – including Random Forest, XGBoost, Logistic Regression, Decision Tree, and Support Vector Machine (SVM) – to predict customer deposit behaviour, XG Boost emerged as the standout performer. It displayed the highest accuracy, earning its place as the chosen algorithm for further refinement and integration into our predictive model. Utilizing libraries like scikit-learn and XGBoost, this selection lays the groundwork for our subsequent in-depth analysis and insights. The detailed accuracy of each algorithm is provided below.

Algorithms	Accuracy
Random Forest	0.8511376568564646
XG Boost	0.8570788584492093
Logistic Regression	0.8018162691085402
Decision Tree	0.7792831392646744
Support Vector Machine (SVM).	0.7205478514275665

TABLE - Comparison Between classification algorithms

The evaluation of the machine learning model's performance in predicting customer deposit behavior involves a detailed examination of the confusion matrix. The matrix, displayed as [[989, 190], [130, 922]], encapsulates the essence of the model's predictive accuracy by categorizing its predictions into true positives (922), true negatives (989), false positives (190), and false negatives (130). True positives represent instances where the model correctly predicted customers making deposits, while true negatives capture accurate predictions of customers not making deposits.

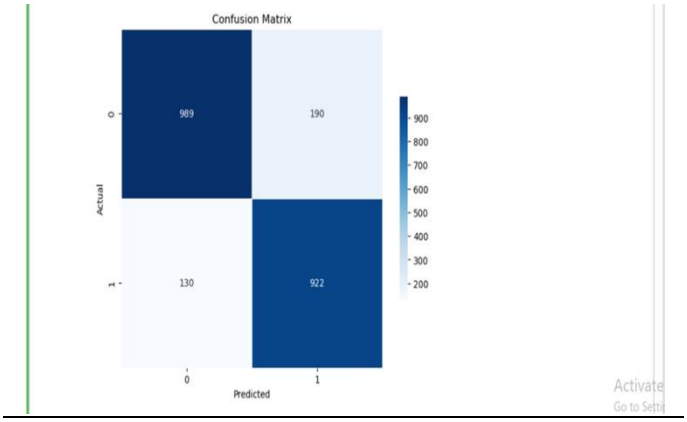

Fig 3: Confusion Matrix

False positives indicate cases where the model erroneously predicted deposits, and false negatives represent instances where it failed to predict actual deposit occurrences. This nuanced breakdown goes beyond conventional accuracy metrics, offering a granular understanding of the model's strengths and weaknesses. Notably, while the model excels in identifying customers making deposits, it exhibits some challenges in predicting non-deposits. This analysis provides valuable insights for refining the model, guiding future iterations, and optimizing its predictive capabilities for improved customer deposit forecasts.

The predictions distribution plot offers a succinct visualization, delineating the count of instances predicted as Deposit and No Deposit. Depicted through two distinct bars, each indicative of the frequency of predictions for the respective classes, the plot provides an immediate overview of the model's performance tendencies. The height of each bar corresponds to the count of predictions, offering insights into the model's ability to identify customers likely to subscribe to term deposits (Deposit Predictions) and those less inclined to do so (No Deposit Predictions). This visual aid serves as a rapid diagnostic tool, enabling a quick assessment of the model's overall performance and potential biases, complementing the nuanced insights gained from numerical metrics.

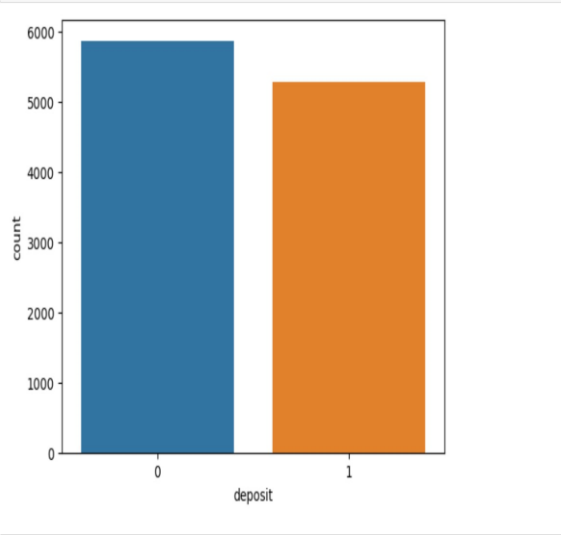

Fig 4: Predictions Distribution Plot

The scatter plot showcased in Figure 2 serves as a visual representation of the predicted probabilities assigned by the model for each data point, conveying the likelihood of making a deposit. Along the x-axis, the continuum of predicted probabilities reveals the model's varying degrees of confidence in predicting deposit outcomes, while the y-axis corresponds to individual data points. Each point is color-coded based on the actual label, with 0 denoting instances of no deposit and 1 indicating instances of deposit. This color-coded distinction allows for an immediate visual assessment of how well the model's predicted probabilities align with the actual outcomes. The scatter plot provides valuable insights into the dispersion and clustering of points, offering a nuanced understanding of the model's calibration and its effectiveness in accurately predicting customer deposit behaviour.

VI. Conclusion

In conclusion, our exhaustive exploration of machine learning tools—Random Forest, XG Boost, Logistic Regression, Decision Tree, and Support Vector Machine (SVM)—for predicting customer deposit behaviour culminated in the unequivocal selection of XG Boost as the preeminent performer. With its distinguished performance, XG Boost emerged as the algorithm of choice for further refinement and integration into our predictive model. The strategic utilization of scikit-learn and XG Boost libraries fortified the foundation for our subsequent in-depth analysis, laying the groundwork for insightful revelations into customer behaviour. This strategic selection not only validates our commitment to precision but also positions our predictive model at the forefront of enhancing strategies in the dynamic banking landscape.

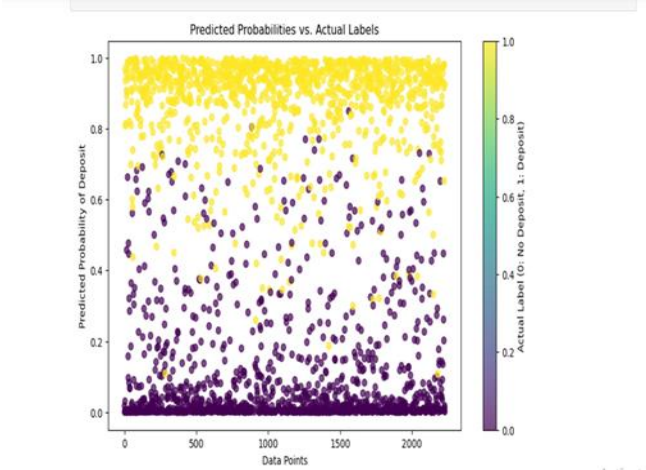

Fig 5:Scatter Plot

VII. Future Work

Looking forward, our exploration into customer deposit behaviour prediction opens avenues for future enhancements. One key avenue lies in the augmentation of features, where a more extensive array could be incorporated to deepen our understanding of customer dynamics. Building upon the success of XG Boost, there exists potential for the integration of more advanced classification models, such as exploring ensemble methods or delving into the intricate landscapes of

neural networks. Future iterations could leverage additional attributes, going beyond the current set, to uncover latent patterns and refine predictive accuracy. The prospect of infusing our predictive models with the prowess of ensemble learning, while considering interpretability and efficiency, stands as a pertinent trajectory. Furthermore, as the industry gravitates towards the transformative realms of Artificial Intelligence, the fusion of machine learning algorithms with AI-driven approaches becomes a compelling avenue. Integrating Deep Learning techniques into our predictive arsenal holds promise for unraveling intricate customer behaviour nuances, ensuring adaptability to the dynamic banking landscape. As we chart this course, the intersection of traditional algorithms, novel features, and the evolving landscape of Deep Learning emerges as the compass guiding our future endeavors in predictive analytics for the banking sector.

REFERENCES

[1] Md. Mehedi Hassan , Eshtiak Ahmed, Mohammad Abu Tareq Rony, Asif Karim ,Sami Azam and D.S.A. Aashiqur Reza “Identifying Long –Term Deposit Customers: A Machine Learning Approach”, IEEE International Conference on Informatics and Software Engineering,2021

[2] 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.

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

[4] 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.

[5] T. Yang, K. Qian, D. C. Lo, Y. Xie, Y. Shi and L. Tao, ”Improve the Prediction Accuracy of Na¨ıve 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 .

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

[7] C. S. T. Koumetio, 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.

[8] A. Jimenez-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.

[9] 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 .

[10] 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.