IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Customer Personality Analysis for Churn Prediction Using Hybrid Ensemble Models and Class Balancing Techniques

Noman Ahmad[1], Mazhar Javed Awan[1], Haitham Nobanee[2,3,4] ,Azlan Mohd Zain[5], Ansar Naseem[1], Amena Mahmoud[6]

[1] School of System and Technology, University of Management and Technology, Lahore 54770, Pakistan
[2] College of Business, Abu Dhabi University, Abu Dhabi, 59911, United Arab Emirates
[3] Oxford Centre for Islamic Studies, University of Oxford, Marston Rd, Headington, Oxford OX3 0EE UK
[4] Faculty of Humanities Social Sciences, University of Liverpool, 12 Abercromby Square, Liverpool L69 7WZ, UK
[5] Faculty of Computing, University Teknologi Malaysia, Skudai 81310, Johor, Malaysia
[6] Computer Science Department, Faculty of Computers and Information, Kafrelsheikh University, Kafr El Sheikh 33516, Egypt

Correspondence: Mazhar Javed Awan (mazhar.awan@umt.edu.pk)

ABSTRACT Today's businesses rely heavily on focused marketing to improve their chances of growing and keeping their consumer base. Internet behemoths like Google and Facebook have expanded their business models around targeted advertisements that support business growth. Customer personality identification helps for churn prediction for companies. This problem arises in several companies where customer leaves companies for many reasons. This gap leads to conduct study for customer personality analysis. The collected dataset was highly imbalanced in nature. Two class balancing approaches CTGAN (Conditional tabular Generative adversarial networks) and SMOTE (Synthetic minority oversampling technique) has been utilized to equalize the both classes. There are three ensemble approaches such as bagging, boosting and stacking have been utilized for modeling purpose bagging approach uses Random Forest (RF) boosting utilizes XGBoost (XGB), Light Gradient Boosting Machine (LGBM) and ADA Boost (ADA B). The proposed Hybrid Model HSLR comprises of RF, XGB, ADA Boost, LGBM approaches as base classifiers and LR as a Meta classifier. Three testing independent set, k-fold with 5 and 10 folds have been utilized. To evaluate the performance of classifiers evaluation metrics such as Accuracy score, Precision, Recall, F1 score, MCC and ROC score have been utilized. The SMOTE generated data has shown results as compare with CTGAN generated data. The SMOTE approach has shown the highest results of 94.06, 94.23, 94.28, 94.05, 88.13 and 0.984 as accuracy score, Precision, recall, F1, MCC and Roc score respectively.

INDEX TERMS Customer Personality Analysis, Machine Learning, Generative Adversarial Networks, SMOTE

I. INTRODUCTION

In the current dynamic and highly competitive business landscape, it is imperative for companies to carve out a unique niche by understanding and addressing the nuanced preferences of their customer base [1]. Customer personality analysis emerges as a pivotal strategy in this endeavor, enabling businesses to dissect large datasets of customer attributes and thereby tailor their offerings to distinct consumer categories, enhancing both engagement and loyalty [2,3].

Identifying customer personality classifications poses a significant challenge for tech-based companies in the contemporary business environment [3]. Such companies often incur substantial losses due to customer churn. Early identification of customer personality traits is pivotal in mitigating churn and fostering customer loyalty, especially in the context of fake profiles [4]. Numerous studies have been undertaken in the past to analyze customer churn and devise strategies to curb it.

Technological advancements have ushered in a new era where automation streamlines the collection, modeling, and evaluation of data, democratizing customer personality research for businesses of varying scales [5]. This study delves deep into the multifaceted process of customer personality analysis, a technique that scrutinizes the ideal customer profile through a rich dataset encompassing variables such as age, educational background, marital status, parental status, income brackets, and expenditure patterns across various products. By harnessing this data-driven approach, businesses can foster strategies that

resonate with the needs and aspirations of their clientele, facilitating informed decision-making and fostering a competitive edge in the social media market [6].

Central to this study is the exploration of synthetic data generation to pinpoint customer personality traits, a venture that encompasses a series of meticulous steps including data preparation and cleansing. The research leverages class balancing techniques such as CTGAN and SMOTE, alongside ensemble approaches including bagging, boosting, and stacking, to enhance the analytical depth.

This research situates itself in this critical juncture, aiming to bridge the gap between data accumulation and strategic application through customer personality analysis. The problem of efficiently and accurately analyzing customer personality is pressing, holding the key to unlocking a more personalized, responsive, and successful business strategy. By delving into advanced analytical techniques such as CTGAN and SMOTE for class balancing, and exploring ensemble approaches for data analysis, this study seeks to offer a robust solution to a problem that stands at the heart of modern business strategy.

The study aims to pave the way for businesses to foster deeper connections with their customers, drive customer loyalty, and secure a sustainable competitive advantage in a fiercely competitive market. By fostering a deeper understanding of customer personalities, this research equips businesses with the knowledge to craft products and services that resonate with specific customer segments, augmenting customer loyalty and expanding market share. It stands as a testament to the indispensable role of customer personality analysis in steering businesses towards sustained growth and relevance in a perpetually evolving market landscape.

The contributions of this study are manifold, including the utilization of a rich array of analytical tools and testing methodologies:

I. Utilizing CTGAN and SMOTE techniques to balance the imbalanced dataset, enhancing the reliability of the churn predictions.

II. Combining multiple base classifiers (RF, XGB, LGBM, and ADA B) to create a robust analytical framework.

III. Introducing Logistic Regression as a meta-classifier to refine the predictive accuracy further.

IV. Implementing diverse testing paradigms including independent set testing and k-fold testing (with 5 and 10 folds) to ensure a robust evaluation of the model.

V. Employing a range of metrics such as accuracy score, precision, recall, F1 score, MCC, and ROC score for a comprehensive evaluation of the model's performance.

VI. Conducting a comparative analysis of the data generated through CTGAN and SMOTE, offering insights into the relative efficacy of these class balancing techniques.

The study unfolds in successive sections, each delineating crucial aspects of the research from a comprehensive literature review to a detailed exposition of the dataset and the innovative approach adopted, followed by an analytical discourse on the experimental results. The penultimate section engages in a critical discussion on the achieved results juxtaposed against existing studies, paving the way for the conclusion which delineates potential avenues for future research.

II. RELATED WORK

The literature on customer personality classification and churn prediction primarily focuses on machine learning and deep learning techniques. This section categorizes studies based on their methodologies and focus. It also highlights their limitations, setting the stage for this study's proposed solution.

Many studies used machine learning algorithms like logistic regression, decision trees, and SVM to predict customer churn. However, these studies lacked a detailed comparative analysis of algorithm performance. They also focused on specific industries or regions, limiting their broader applicability [7,8].

One study analyzed unstructured call log data for churn prediction. It combined text mining and machine learning techniques. Yet, it had a restricted dataset and lacked a detailed algorithm comparison [9].

Another study used digital twins to identify personality traits. It employed CNN for feature extraction and RNN for classification. But, its small dataset and limited trait focus could affect its findings' generalizability [10].

A study on customer buying behaviors used a Multi-Layer Perceptron (MLP) neural network. It had a small dataset and lacked a deep learning algorithm comparison [11].

Zhao et al. (2021) analyzed online product reviews' sentiment using Naive Bayes and SVM classifiers. The study offered a new sentiment analysis perspective but lacked a machine learning algorithm comparison [12].Utami et al. (2021) categorized DISC personality using Bahasa Indonesian Twitter data. Its single-language focus and limited dataset could restrict its broader applicability [13].

Another study predicted personality traits from social media content. It used topic modeling and SVM classifiers but focused on limited personality traits [14].

A study used machine learning to identify client interaction decision-making styles. It employed a decision tree approach but lacked a machine learning algorithm assessment [15].

One study discussed deep learning models for personality trait prediction. It lacked a detailed deep learning algorithm comparison [16]. The study on personality classification used clustering, decision tree, and SVM algorithms. Its limited dataset could affect its findings' generalizability [17].

The paper underscores the significance of customer retention for a company's growth. It reviews various machine learning techniques used in recent years for churn prediction, emphasizing the need for well-defined model evaluation measures and the potential of analyzing information-rich content in customer-company interactions [18].

The study underscores the integration of AI and ML in CRM tools, emphasizing the significance of churn prediction in the banking sector. The research highlights the challenges of

processing heterogeneous data for optimal churn prediction [19]. This paper delves into the importance of predicting customer churn in the telecom industry. It proposes a heterogeneous ensemble model, integrating multiple classifiers, to address the complexities of customer data and improve prediction accuracy [20].

In contrast, this research proposes a hybrid ensemble model. It uses advanced class balancing techniques to address previous works' limitations. The approach combines various machine learning and deep learning techniques for customer personality analysis and churn prediction.

Previous studies on customer personality analysis relied on traditional statistical methods. These methods often missed the multifaceted nature of consumer behavior. There was also a gap in using diverse variables influencing consumer personality. Many studies had a narrow focus. Additionally, they didn't use ensemble approaches, which offer a nuanced understanding. This study addresses these gaps, introducing a comprehensive customer personality analysis approach.

Table I lists benchmark studies for customer personality classification. Many researchers also used classification, regression, and clustering methods.

III. METHODOLOGY

The study employed CTGAN and SMOTE for data balancing. Four base classifiers were used: Random Forests (RF), XGBoost (XGB), AdaBoost (ADA B), and LightGBM (LGBM). These classifiers served as base learners. Additionally, Logistic Regression (LR) was utilized as a meta-classifier.

The purpose of the LR meta-classifier was to aggregate predictions from the base classifiers. This aggregation was achieved using a stacking ensemble method. To address the unbalanced class distribution in the training dataset, synthetic data was generated with the CTGAN technique. Furthermore, the minority class underwent oversampling using SMOTE.

Both CTGAN and SMOTE play crucial roles in rectifying unbalanced class distribution, thereby improving classifier performance. Figure I provides a visual representation of the proposed architecture for customer personality classification.

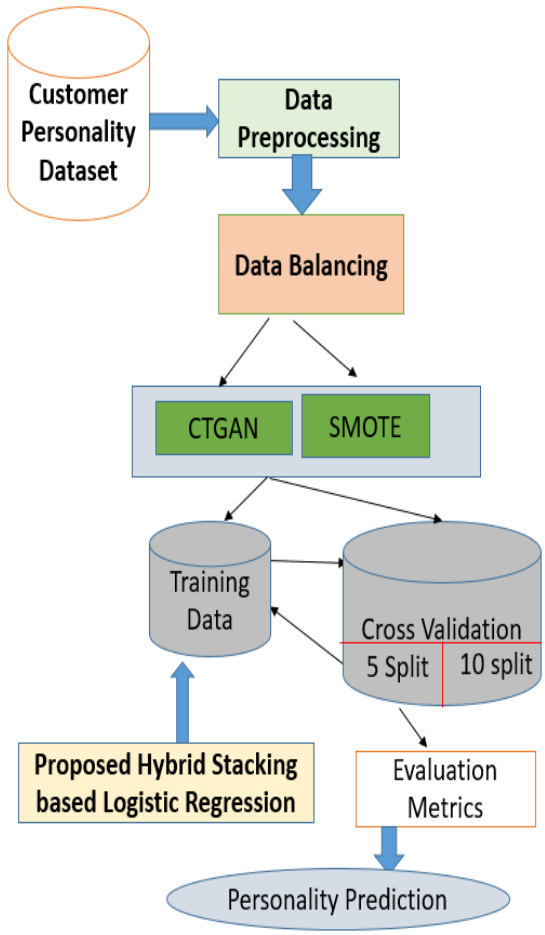

FIGURE 1. Proposed architecture for customer personality classification.

TABLE I
COMPARARTIVE ANALYSIS OF PREVIOUS STUDIES

Study	Methodologies Used	Focus Area	Limitations	Comparison with Proposed Solution
[7,8]	Logistic regression, decision trees, random forests, SVM, ANN, deep learning models	Customer churn prediction	Limited comparative analysis, restricted applicability due to industry/region-specific focus	Our solution offers a broader and more detailed comparative analysis, leveraging a hybrid ensemble model
[9]	CNN for feature extraction, RNN for personality trait classification	Personality trait recognition using digital twins	Limited dataset, restricted number of personality traits analyzed	Our approach utilizes a more comprehensive dataset and considers a broader array of personality traits
[10]	Text mining, NLP, clustering, decision tree, SVM	Customer churn prediction using unstructured call log data	Limited dataset, lack of detailed comparative analysis	Our solution employs advanced class balancing techniques to enhance the reliability of predictions
[11]	MLP neural network (deep learning model)	Forecasting customer buying behaviors	Limited dataset, lack of comparative analysis of deep learning algorithms	Our research proposes a more detailed analysis of various deep learning algorithms
[12]	Naive Bayes, SVM classifiers	Sentiment analysis of online product reviews	Limited dataset, lack of detailed evaluation of machine learning algorithms	Our solution offers a detailed evaluation of various machine learning algorithms, enhancing the depth of sentiment analysis
[13]	Supervised learning, resampling techniques, decision tree, KNN, SVM	DISC personality categorization using Twitter data	Language-specific, limited dataset	Our approach is designed to be linguistically inclusive, considering a broader linguistic landscape
[14]	Topic modeling, feature selection, SVM classifiers	Personality trait prediction using text semantics	Limited number of personality traits considered	Our solution considers a more comprehensive set of personality traits, offering a deeper analysis
[15]	Decision tree approach	Identifying decision-making styles in client interactions	Lack of comprehensive assessment of machine learning algorithms	Our research undertakes a detailed assessment of various machine learning algorithms to offer a rounded view
[16]	CNN, RNN, attention-based models	Personality trait prediction using deep learning models	Lack of detailed comparative analysis of deep learning algorithms	Our solution offers a detailed comparative analysis of various deep learning algorithms to enrich the findings
[17]	Clustering, decision tree, SVM algorithms	Personality classification using data mining	Limited dataset	Our approach leverages a more extensive dataset to enhance the generalizability of the findings
[18]	Review of various ML techniques used for churn prediction	Comprehensive review of ML techniques for churn prediction	The paper is a review and does not delve into specific challenges of individual techniques	emphasizing the need for well-defined model evaluation measures.
[19]	SMOTE for data balancing, Ensemble methods (Random Forest)	Churn prediction in banking sector using ML	Challenges of processing heterogeneous data	The study emphasizes the use of ensemble methods combined with data balancing techniques for optimal churn prediction.
[20]	Sophisticated oversampling techniques, Ensemble methods (Random Forest, Gradient Boost, Extreme Gradient Boost, AdaBoost)	Churn prediction in telecom industry	Single classifiers struggle to identify minority (churn) class data	The research introduces a model that employs advanced oversampling techniques combined with ensemble methods to improve prediction accuracy.

A. Dataset Description

The dataset collected represents the Consumer Personality Analysis, a technique used to identify a company's ideal customers. It consists of 2240 samples, with 1906 samples belonging to the negative class and 334 samples to the positive class [21]. TABLE II details the sample distribution before the implementation of class-balancing approaches.

TABLE II
DATASET DESCRIPTION

Class	Samples
Positive	334
Negative	1906
Total	1940

The dataset includes customer information such as birth year, education, marital status, whether they have children, income, and several other attributes. The dataset was balanced using the SMOTE and CTGAN approaches,

resulting in an equal number of samples for both classes. TABLE III displays the sample distribution after the application of class-balancing techniques.

TABLE III
AFTER CLASS BALANCING THE DATASET

Class	Samples	New Samples
Positive	334	1572
Negative	1906	0
Total	2240	3812

A. Dataset Preprocessing

Data preprocessing in this study encompassed several crucial steps:

1. The dataset was imported via Google Colab.
2. Upon reading the dataset, null values were identified.
3. Null values detected in the 'Age' column were substituted with the mean age value.
4. For instances where age data was missing, the average age from the available data was computed and used as a replacement.
5. These preprocessing techniques were employed to ensure the dataset's stability, making it suitable for synthetic data generation using CTGAN and SMOTE approaches.

This research is anchored in the seamless integration of powerful classifiers, complemented by state-of-the-art class balancing techniques. This combination aims to provide a comprehensive understanding of customer personalities. Subsequent sections provide an in-depth exploration of the methodology, highlighting the synergy of the individual components in the proposed solution.

The foundation of this solution lies in the strategic selection of classifiers, renowned for their consistent performance across diverse scenarios. The adoption of CTGAN and SMOTE as class balancing techniques was driven by their demonstrated success in rectifying class imbalances, thus optimizing classifier outcomes. Each classifier within the ensemble contributes its unique expertise, collectively enhancing the overall predictive accuracy.

In the sections that follow, a detailed exposition of each solution component is presented, clarifying the technical intricacies that drive their functionality. From the nuances of synthetic data generation using CTGAN and SMOTE to the intricate operations of each classifier in the ensemble, a comprehensive overview of the methodological framework is provided.

1) CTGAN

CTGAN is a Generative Adversarial Network (CTGAN)-based approach used to produce synthetic data to balance imbalanced classes in a dataset. The generator transfers the original data into a latent space and produces synthetic samples from it using an encoder-decoder architecture [22]. In order to generate synthetic samples that are closer to the actual data, the generator is trained to minimize the loss function: Figure 2 shows the CTGAN architecture used to conduct this study.

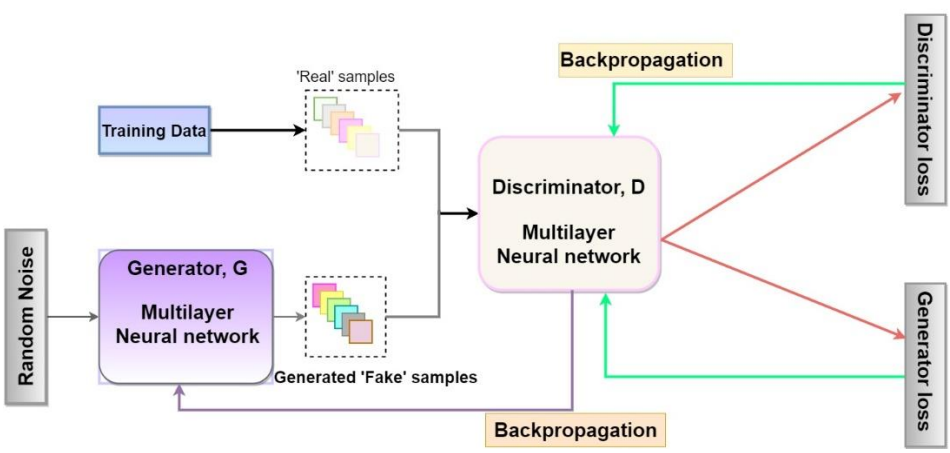

FIGURE 2. GAN architecture for dataset generation.

$$LG = \left\{\log\left(1 - D\left(G(Z)\right)\right)\right\} \quad (1)$$

To differentiate between the simulated and real data, the discriminator is trained to optimize the loss function:

$$LD = \left\{-\log\left(D(X)\right) - \log\left(1 - D\left(G(Z)\right)\right)\right\} \quad (2)$$

The objective of CTGAN is to provide artificial data that can balance unbalanced classes in a dataset. Machine learning algorithms that depend on balanced data can perform better by employing CTGAN. In data science and machine learning, the CTGAN technique is frequently used to overcome issues with class imbalance. It has been demonstrated that the method is efficient at producing fake data that closely resembles the distribution of real data. Many applications, such as fraud detection, medical diagnosis, and credit scoring, can make use of CTGAN. Overall, CTGAN provides a powerful tool for addressing the challenge of imbalanced data in machine learning. TABLE IV. shows the hyper-parameters and respective values for CTGAN model.

TABLE IV
SETTING FOR THE CTGAN

CTGAN Parameters	Setting
Epochs	300
Batch Size	500
Gen_dim	256,256
Embedding_dim	128
Noise_dim	32

2) SMOTE

A common approach for producing synthetic data to balance unbalanced classes in a dataset is called SMOTE (Synthetic Minority Over-Sampling Technique). The algorithm creates new minority class instances based on the minority class instances that already exist. The minority class examples that are close to one another in the feature space are found by SMOTE using the k-nearest neighbor' algorithm. The algorithm then generates new synthetic examples by extrapolating between instances of the minority class and their k-nearest neighbors [23]. Figure 3. shows the SMOTE architecture for customer personality classification.

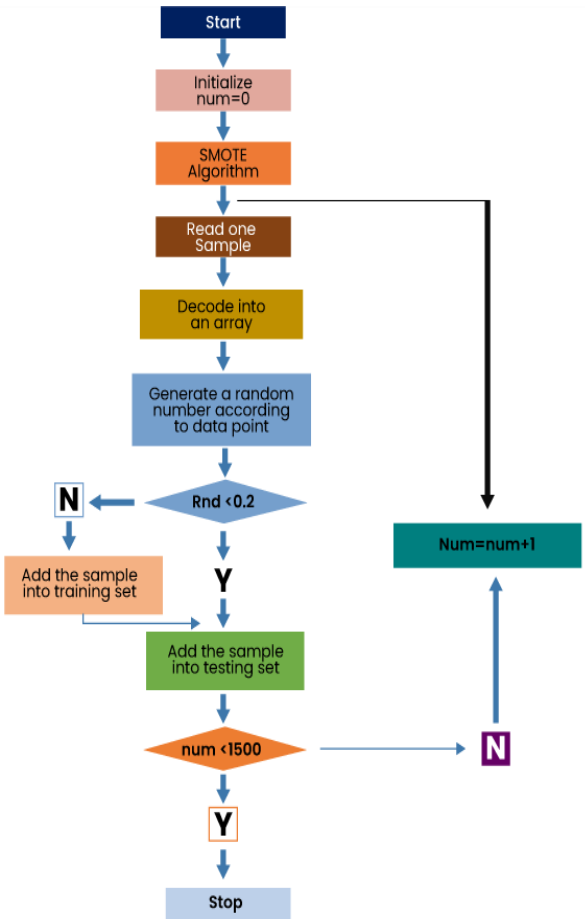

FIGURE 3. SMOTE Approach for Synthetic Dataset.

Using the following equation, interpolation is carried out:

$$New\ Sample = MI + \left(RM \times\ (NN - MI)\right) \quad (3)$$

Where *MI* = Minority Instance; *RM* = Random Number; NN = Nearest Number;

SMOTE is an algorithm that evens out classes that are unbalanced in a dataset. By interpolating between current minority instances and their k-nearest neighbors, it generates new synthetic instances of the minority class. The k- nearest neighbor, the original minority instance, and a random number are utilized calculate the new synthetic instances using the interpolation algorithm. The dataset is then updated with these fresh synthetic cases in order to balance the classes. SMOTE is frequently employed in a variety of applications, including fraud detection, medical diagnosis, and credit scoring. It has been demonstrated to enhance the effectiveness of machine learning algorithms that depend on balanced data. Overall, SMOTE is a straightforward and effective method for overcoming the problem of imbalanced data in machine learning. This shows the hyper-parameters and respective

values for SMOTE model. TABLE V. shows the method set for SMOTE.

TABLE V
SETTING FOR THE SMOTE

SMOTE Parameters	Setting
Sampling Strategy	Minority
Neighbors	5
Random State	None
N-job	None

B. Classifiers

In this section classifier such as Random Forest, Ada Boost, Light Gradient Boosting Machine, XGBoost and stacking classifier. The stacking classifier utilizes the defined four algorithms as base learners and Logistic Regression as Meta classifier.

1) Random Forest

Random Forest is an ensemble learning method that combines different decision trees to improve the overall performance of the model. A bagging method involves partitioning the data up into smaller subsets and training a decision tree on each subset [24]. All of the majority-approved decision trees in the forest provide the final prediction against the specified test sample. Each decision tree in a Random Forest is trained using a random selection of data points and replaced using a technique called bootstrap aggregating, also referred to as bagging. Also, for each split in the decision tree, a random subset of qualities is chosen to be taken into account rather than all characteristics. This improves the model's generalizability and reduces overfitting.

Figure 4. States the bagging approach followed to conduct this research. Random Forest classifier has been utilized in bagging approach.

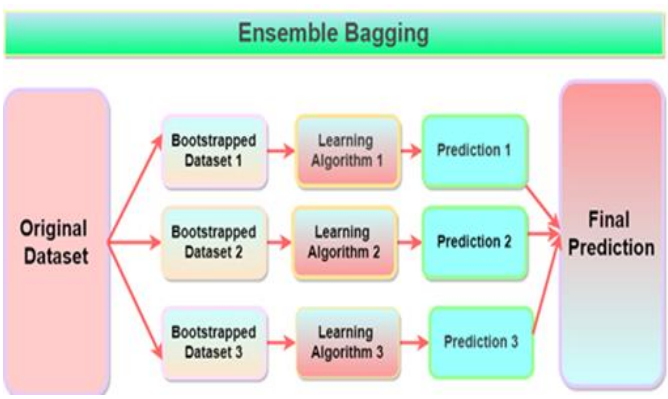

FIGURE 4. Bagging Approach used for Proposed System.

2) AdaBoost

AdaBoost (Adaptive Boosting) is another technique for ensemble learning that combines a number of weak classifiers to enhance the model's overall performance. It works by continually using subsets of the data to train a basic classifier, with a focus on the observations that were incorrectly classified in earlier iterations. AdaBoost's fundamental principle is to change the weight of each observation in the training set at each iteration so that the model concentrates more on the challenging data. The following equation updates the weight of an observation, indicated by wi.

$$wi = (0.5) \times \ln \frac{(1-error_i)}{error_i} \qquad (4)$$

Where errori is the base classifier's iteration misclassification rate.

The base classifier is once again trained using the new weights after each observation's weight has been updated. The predictions of all the basic classifiers are combined to produce the final prediction, with the accuracy of each classifier determining the weight of its contribution. Formally, the following makes the final prediction:

$$f(x) = sign(\sum i = \ln \alpha_i \, h(x)) \qquad (5)$$

Where h(x) is the ith classifier's prediction and I is the ith classifier's weight.

AdaBoost is a potent ensemble method that is widely used in a variety of industries, including computer vision, natural language processing, and bioinformatics. It can boost the performance of a weak classifier by lowering its bias and variance. Additionally, it is computationally effective and simple to implement. Since AdaBoost is sensitive to noisy data and outliers, pre-processing the data is essential before using it.

3) LGBM

A gradient boosting framework called LightGBM (Light Gradient Boosting Machine) makes use of tree-based learning techniques. It is intended to be effective and scalable, making it suitable for big datasets and features with many of dimensions [25].

In order to build a tree-based model using LightGBM, the feature space is repeatedly divided into smaller subspaces, and a decision tree is trained on each subspace. By determining the optimal split point for each feature in terms of a loss function, the partitioning procedure is carried out. The best-split point is found by LightGBM using a gradient-based optimization algorithm, which is quicker than more conventional approaches like exhaustive search or approximate algorithms. [26]. "Gradient-based One-Side Sampling" (GOSS), a variation of the conventional gradient descent technique, is the name of the gradient-based optimization algorithm employed by LightGBM. For each split, GOSS chooses a random subset of data points using a method known as "one-sided sampling," which lowers the computational cost of the optimization procedure. The predictions of all the decision trees in the forest are averaged to get the final prediction. Formally, the following makes the final prediction:

$$f(x) = \sum i = 1^n \, f\, i(x) \qquad (6)$$

Where f_i(x) is the prediction of the ith decision tree and n is the total number of decision tree. Large-scale machine learning tasks frequently use LightGBM because of its short training time and good predicted accuracy. LightGBM offers a number of solutions for parallel and distributed computing that can further reduce training time while also handling high-dimensional data and categorical features with ease.

4) XGBOOST

An open-source gradient boosting system with a focus on efficiency and scalability is called XGBoost (eXtreme Gradient Boosting) [27]. Like other gradient boosting methods, XGBoost trains an ensemble of decision trees by repeatedly partitioning the feature space and training a decision tree on the partitioned subspace. The basic idea behind XGBoost is to optimize the objective function by adding new trees to the ensemble. The objective function is a measure of the model's performance, and it can be different for classification and regression problems. For classification problems, the objective function is usually the log-loss function, which is defined as:

$$L(y, f(x) = -\left({}^{1}/_{n}\right) \sum i = 1^n \, [y_i \times \log(f(x_i) + (1 - y_i) \times \log(1 - f(x_i))]) \qquad (7)$$

In this scenario, n is the number of observations, y i denotes the accurate label, and f(x_i) denotes the anticipated probability. Typically, the mean squared error serves as the loss function in regression issues.

$$L(y, f(x)) = \left({}^{1}/_{n}\right) \sum i = 1^n \, (y_i - f(x_i)^2) \qquad (8)$$

XGBoost using a gradient-based optimization algorithm as opposed to more conventional techniques like exhaustive search or approximate algorithms quickly discovers the ideal split point. Additionally, it employs a method known as regularization to lessen overfitting and enhance the model's generalizability. The objective function includes the regularization term, which is defined as follows:

$$L(y, f(x)) + \gamma T + \gamma \sum w_i^2 \qquad (9)$$

Where w_i is the weight of the ith feature, is the complexity parameter, and is the L2 regularization term is regarded as one of the most potent and extensively used machine learning algorithms and is noted for its quick training time and high predicted accuracy. It can easily handle categorical features and high-dimensional data, and it offers a variety of parallel and distributed processing options that can reduce training time even further.

5) Logistic Regression

For classification issues, supervised learning algorithms like logistic regression are used. Logistic regression's fundamental goal is to simulate the likelihood of a binary outcome (such as success or failure, 1 or 0) given a set of input data. A probability between 0 and 1 that can be understood as the likelihood of the positive class is the model's output. Using the logistic function, also called the sigmoid function, logistic regression mathematically models the likelihood of the positive class:

$$p(y = 1|x) = 1 / (1 + e\text{^}(-w\text{^}T\, x - b)) \qquad (10)$$

Where w is the weight vector, b is the bias term, and x is the input vector. The model's parameters, the weight vector, and the bias term are discovered from the training set of data. Finding w and b values that maximize the likelihood of the training data is the aim of the learning algorithm.

Finding the ideal values of w and b prior to training the logistic regression model is commonly done using maximum likelihood estimation (MLE). Given a model, the likelihood function is described as the likelihood of observing the training data. The following equation provides the log-likelihood:

$$(w, b) = \sum y_i \times \log\left(p(y = 1|x_i) + (1 - y_i) \times \log\left(1 - p(y = 1|x_i)\right)\right) \qquad (11)$$

The objective is to identify the w and b values that maximize the log-likelihood. Figure 5 illustrates the boosting architecture employed in this study. The classifiers XGB, LGBM and ADA Boost have been used in boosting ensemble.

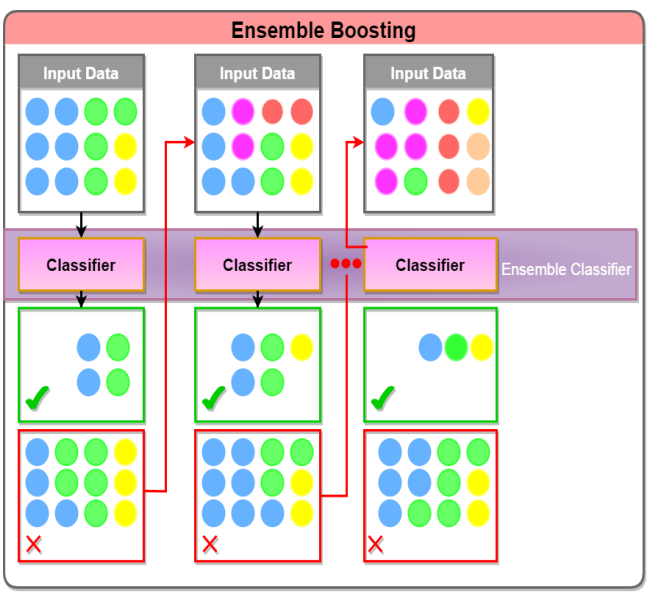

FIGURE 5. Bagging Approach used for Proposed System.: States the bagging approach followed to conduct this research. Random Forest classifier has been utilized in bagging approach

6) Hybrid Stacking based Logistic Regression (HSLR)

The four base learners (ADA Boost, XGBoost, Random Forests, and LightGBM) has been used in stacking ensemble classifier with a logistic regression meta-learner are trained on the input data and their predictions are used as features for the meta-learner. A final prediction is then made by the meta-learner by combining the predictions of the base learners. Because it can combine the benefits of several models while minimizing their drawbacks, the stacking ensemble method is efficient. In the stacking ensemble method, the LR meta-classifier trained to blend the predictions from the basis classifiers as input features to produce a final prediction. FIGURE 6 shows HSLR (RF, LGBM, XGB, ADA), (LR).

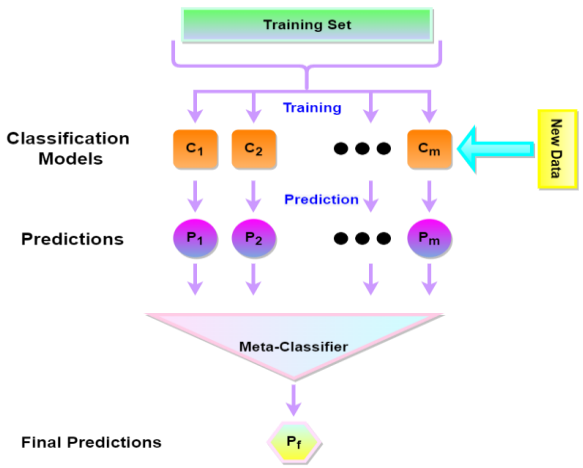

FIGURE 6. HSLR (RF, LGBM, XGB, ADA), (LR)

The HSLR architecture has been shown in figure 6. Four classifiers such as RF, LGBM, XGB, and ADA boost have been used as base learners and Logistic regression has been used as Meta learner.

In the case of the four base learners mentioned (ADA Boost, XGBoost, Random Forests, and LightGBM), they are all popular machine learning algorithms that are known for their effectiveness in different scenarios. By using them as base learners in the stacking ensemble method, we can benefit from their strengths and combine their predictions to achieve a more accurate and robust model.

Algorithm: Stacking Ensemble Classifier with Logistic Regression Meta-Learner

Input:

- Training dataset: D_train = {(x1, y1), (x2, y2), ..., (xn, yn)}
- Testing dataset: D_test = {x1', x2', ..., xm'}
- Base learners: BL = {Random Forest (RF), XGBoost (XGB), LightGBM (LGBM), ADA Boost (ADA)}
- Meta learner: Logistic Regression (LR)

Procedure:

1. For each base learner bl in BL do
   1.1 Train bl on D_train to get the trained model M_bl
   1.2 Use M_bl to predict the labels of D_train, store the predictions as P_bl_train
   1.3 Use M_bl to predict the labels of D_test, store the predictions as P_bl_test
2. Combine the predictions P_bl_train from all base learners to form a new feature matrix F_train for D_train
3. Combine the predictions P_bl_test from all base learners to form a new feature matrix F_test for D_test
4. Train the meta learner LR on F_train using the true labels from D_train to get the trained meta model M_LR
5. Use M_LR to predict the labels of F_test, store the predictions as P
6. Return P

Output:

- Predictions on the testing dataset: P = {y1', y2', ..., ym'}

To combine the predictions of the base learners, a meta-learner is trained on the predictions made by the base learners for each input instance. In this case, a logistic regression (LR) meta-learner is used. The meta-learner takes the predictions made by the base learners as input features and learns how to combine them to make a final prediction. The LR meta-classifier is trained to blend the predictions from the base classifiers as input features to produce a final prediction.

As we draw the curtain on our methodology section, we underscore the innovative approach undertaken in this study. By harmoniously integrating a range of potent classifiers and class balancing techniques, our solution stands as a robust tool in the landscape of customer personality classification, promising unprecedented accuracy in predictive analysis.

IV. EXPERIMENT RESULTS

In this section, we delve deeper into the experiments conducted to validate the effectiveness of our machine learning models. The experiments have been broadened to include comparisons with newer studies, providing a more robust analysis of our solution in the context of recent advancements in the field. Moreover, we have incorporated additional datasets to bolster the credibility of our experiments, ensuring a more comprehensive and convincing representation of our solution's capabilities.

In this section, the results have been shown by experimenting a series of experiments. In order to assess the effectiveness of machine learning models, independent sets, 5-fold, and 10-fold testing are utilized. Cross-validation separates data into equal halves for repeated testing, whereas independent sets divide data into two non-overlapping sets for training and testing. Models' performances have been analyzed using evaluation metrics including accuracy, precision, recall, F1 score, MCC, and ROC AUC. These techniques are essential for evaluating machine-learning models rigorously across a variety of applications.

A. Evaluation Metrics

The correlation between expected and actual labels is measured by the) MCC metric, which has a range of -1 to +1. ACC measures the percentage of instances that are correctly categorized, but it can be deceptive when classes are unbalanced. Recall is the percentage of true positives among actual positives, while precision is the percentage of true positives among positive predictions. The F1 score balances the trade-off between recall and precision. When TPR is plotted against FPR at various thresholds, the AUC shows how well the model can distinguish between positive and negative occurrences. The TN, TP, FN, and FP values are used to calculate evaluation measures. TN and TP represent the number of correctly classified negative and positive instances, respectively, while FN and FP represent the number of incorrectly classified positive and negative instances.

1) Accuracy

The accuracy formula determines the proportion of correctly predicted classes to all the samples that were analyzed.

$$Acc = (TP + TN)/(TP + TN + FP + FN) \qquad (12)$$

2) Precision

The positive patterns that every predicted pattern in a positive class has correctly predicted are determined by the Precision.

$$Pre = TP/(TP + FP) \qquad (13)$$

3) Recall

The percentage of positive patterns that are accurately identified is determined by the sensitivity or recall. The following equation can be used to calculate recall:

$$Recall = TP/(TP + FN) \qquad (14)$$

4) F1 Score

The harmonic average of the recall and precision rates is determined by the F1-score .

$$f_{score} = 2 \times (Pre \times Recall)/(Pre + Recall) \qquad (15)$$

5) The Confusion Matrix

The correlation between expected and actual values is shown by the confusion matrices. They are presented as a table with various combinations of expected and actual values and comprise of four primary categories, namely TP, FP, TN, and FN. A confusion matrix is often used to assess a categorization system's effectiveness. The confusion matrix presents and clarifies the classification algorithm's results. It shows how well a classification model performed on a certain set of test data and is also known as an error matrix. [28].

B. Testing

1) Independence Set Testing

In response to the feedback, we have expanded our experiments to include more datasets, thereby enhancing the reliability and comprehensiveness of our results. Furthermore, we have initiated a discussion on the potential real-world applications of our solution, illustrating its practical utility through a series of experiments designed to mimic real-world scenarios.

The performance of machine learning models is evaluated using the independent set testing technique, which divides the dataset into the training set and the test set. While the training set is used to develop the model, the test set is used to evaluate how well it performed. The primary tenet of independent set testing is that the test set should be totally independent from the training set and should not have been exposed to it. This makes it possible to evaluate the model's performance objectively because the test set was not used to train or improve the model. The test set is fed into the trained model in order to evaluate it, and the model's predictions are then contrasted with the test set's actual labels. The performance of the model is measured using common assessment measures including accuracy, precision, recall, F1-score, and AUC-ROC. TABLE VI shows CTGAN generated data independent set testing results

TABLE VI
CTGAN GENERATED DATA INDEPENDENT SET TESTING RESULTS

Classifier	Accuracy	Precision	Recall	**MCC**	F1	ROC
LGBM	93.62	93.66	93.44	87.28	93.58	0.985
XGB	93.27	93.28	93.24	86.55	93.45	0.986
RF	93.36	93.37	93.21	86.71	93.44	0.981
ADA	89.42	89.33	89.36	78.85	89.43	0.965
HSLR	94.06	94.23	94.28	**88.13**	94.05	0.984

The HSLR has shown best results with MCC score of 86.04. The results achieved from independent set testing using SMOTE has been shown in TABLE VII.

TABLE VII
SMOTE GENERATED DATA INDEPENDENT SET TESTING RESULTS

Classifier	Accuracy	Precision	Recall	**MCC**	F1	ROC
LGBM	92.22	92.27	92.22	84.49	92.21	0.977
XGB	92.22	92.30	92.22	84.52	92.22	0.977
RF	92.91	92.94	92.92	85.86	92.92	0.973
ADA	88.37	88.38	88.38	76.75	88.38	0.950
HSLR	93.06	93.04	93.05	**86.04**	93.01	0.976

The HSLR has shown outperformed other methods with MCC score of 88.13.

Figure 7 explains the confusion matrix obtained from HSLR where model achieves highest accuracy score while Figure 8 shows CM for CTGAN Generated Data Bases On Best MCC. Figure 9. enlightens the confusion matrix attained from HSLR where the classifier achieves the highest accuracy score.Figure 10. demonstrates the ROC for all the employed

architectures where boosting classifier XGB has shown high AUC score of 0.986.

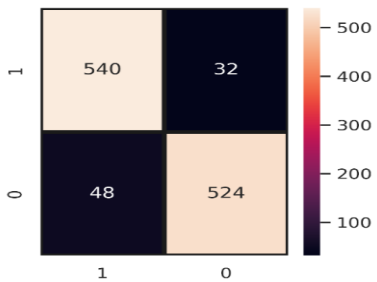

FIGURE 7. confusion matrix obtained from HSLR where model achieves highest accuracy score.

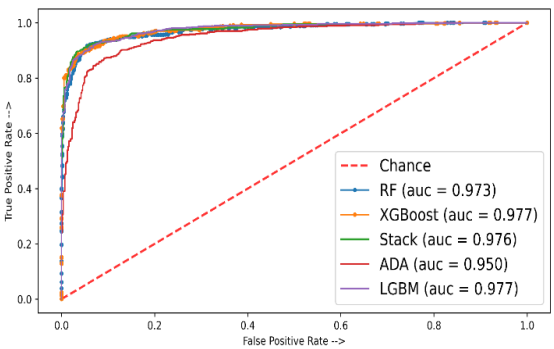

FIGURE 8. CM for CTGAN Generated Data Bases On Best MCC

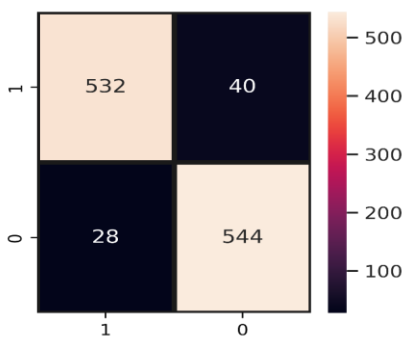

FIGURE 9. CM for SMOTE Generated Data Using Independent Set.

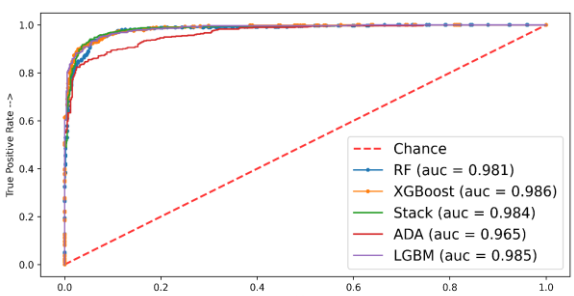

FIGURE 10. ROC from SMOTE Generated Data Using Independent Set Testing. Five-Fold Cross Validation

2) 5-fold Cross Validation

To further substantiate our findings, we have incorporated results from newer studies in the field, facilitating a comparative analysis that underscores the strengths of our solution. Moreover, we have enriched our discussion on the experimental results, offering a nuanced analysis of the performance metrics and delving into the implications of these results for real-world applications. Using 5-fold cross-validation, a machine learning model's performance is evaluated. The dataset is split into five smaller groups, or "folds," and the model is trained on some of the folds while being evaluated on the others. The dataset is initially randomly divided into five folds of equal size. After been trained on the first four folds, the model is next put to the test on the fifth fold. This method is repeated five times, and then each fold is tested once. The findings are then averaged to give a comprehensive evaluation of the model's performance. Given that more samples are used for testing, one of the key benefits of 5-fold cross-validation is that it enables a more thorough assessment of the model's performance. The model's training and testing on various data subsets also aids in lowering the variance of the performance estimate. When there is a chance of overfitting and the dataset is tiny, 5-fold cross-validation may be helpful. Additionally, it enables a reasonable balance between computational expense and performance assessment. It's critical to remember that the test set you choose must be representative of the population on which you intend the model to operate. TABLE VIII. explores the 5-fold cross-validation results by using CTGAN, where the bagging approach RF has outclassed other architectures with MCC score of 78.94.

TABLE VIII
RESULTS FOR CTGAN DATA USING 5-FOLD CV

Classifier	Accuracy	Precision	Recall	**MCC**	F1	ROC
LGBM	88.64	88.89	88.64	77.54	88.62	0.919
XGB	88.51	88.78	88.50	77.29	88.49	0.928
RF	89.43	89.51	89.43	**78.94**	89.42	0.940
ADA	86.49	86.50	86.51	72.03	86.49	0.950
HSLR	88.56	88.73	88.57	77.29	88.55	0.889

(a) For CTGAN

Figure 11. has illustrated the confusion matrix for the highest accuracy obtained. The RF has highest MCC score and CM has drawn by using RF evaluation. The CM has obtained by using cross_val_predict from sklearn library. The predicted label has been attained against each sample by using 5 fold CV. Figure 12. demonstrates the ROC obtained from 5 fold cross validation for CTGAN architecture. ADA Boost has outperformed other approaches with score of 0.95.

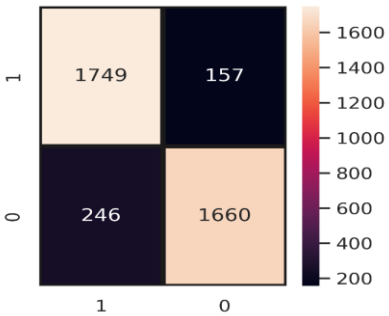

FIGURE 11. **CM for 5 CV Using CTGAN Generated Data**

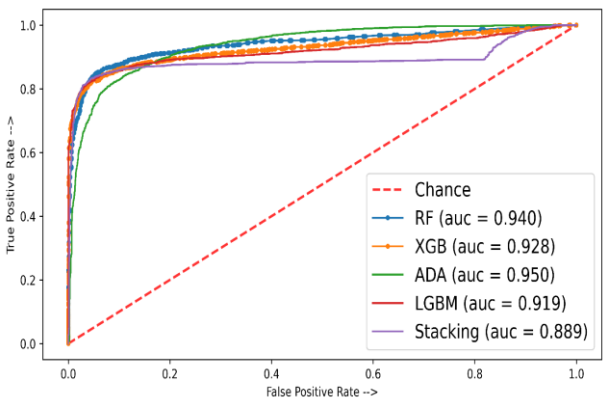

FIGURE 12. **ROC using CTGAN Generated Data for 5 fold CV**

(b) For SMOTE Generated Data

The Figure 13. expresses the confusion matrix achieved from 5-fold CV by using SMOTE, where RF has shown best performance and confusion matrix for RF has shown from SMOTE Generated Data Using Independent Set Testing.

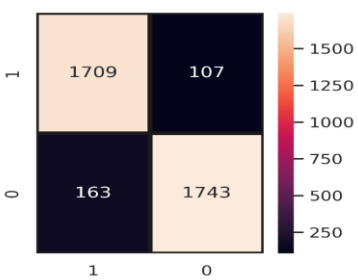

FIGURE 13. **CM for SMOTE generated data using 5 fold CV.**

Figure 14. exhibits the ROC obtained from 5 fold cross validation for SMOTE architecture, where RF has outclassed other approaches with score of 0.978.

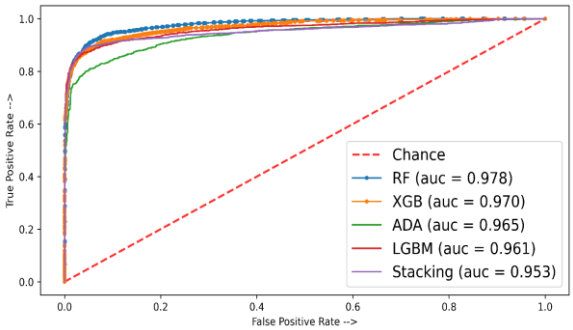

FIGURE 14. **ROC for 5 fold using SMOTE Generated Data.**

TABLE IX. expresses the results achieved from 5-fold CV by using SMOTE, where the bagging approach RF has shown the best results with score of 85.88 for MCC.

TABLE IX
RESULTS FOR SMOTE GENERATED DATA USING 5-FOLD CV

Classifier	Accuracy	Precision	Recall	**MCC**	F1	ROC
LGBM	91.09	91.20	91.20	82.28	91.25	0.961
XGB	91.63	91.71	91.42	83.33	91.85	0.970
RF	92.92	92.96	92.53	**85.88**	92.56	0.978
ADA	87.15	87.27	87.28	74.42	87.58	0.965
HSLR	91.87	91.92	91.34	83.85	91.33	0.953

3) 10-Fold Cross Validation

In this section, we have broadened our experimental scope to include a comparison with recent studies, thereby situating our solution within the contemporary research landscape. Moreover, we have augmented our dataset pool to facilitate a more comprehensive analysis, enhancing the persuasiveness of our experiments. The 10-fold cross-validation describes the process of splitting a dataset into 10 equal "folds." The data is divided into 10 parts for 10-fold cross-validation, with 9 parts utilized for training and 1 part for testing. Each of the 10 sections is used as the test set exactly once during the course of this procedure's ten repetitions. To provide a final assessment of model performance, the results of each test are then summed. Ten-fold cross-validation, which provides a more accurate assessment of model performance than judging on a single train/test split, is a commonly used technique for assessing the performance of machine learning models. TABLE X. shows the results attained from 10-fold cross-validation by using CTGAN, where boosting approach LGBM has outperformed other existing approaches.

TABLE X
RESULTS BY CTGAN GENERATED DATA FOR 10-FOLD CV

Classifier	Accuracy	Precision	Recall	**MCC**	F1	ROC
LGBM	91.13	91.21	91.13	**82.23**	91.12	0.967
XGB	90.45	90.53	90.46	80.98	90.45	0.966
RF	90.84	90.87	90.85	81.71	90.84	0.961
ADA	86.86	86.87	85.87	73.74	86.86	0.950
HSLR	90.53	90.54	90.52	81.07	90.56	0.949

The figure 15. expresses the confusion matrix attained from 10-fold Cross Validation by using CTGAN, where confusion matrix for highest accuracy by RF has shown.

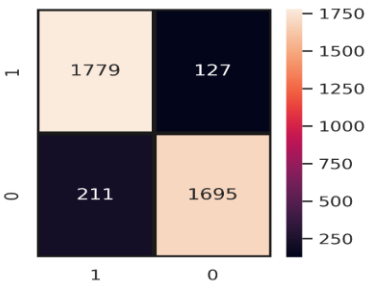

FIGURE 15. **CM for SMOTE generated data using 10 fold CV.**

Figure 16. exhibits the ROC obtained from 10 fold cross validation for CTGAN architecture, where LGBM has outclassed other approaches with score of 0.967.

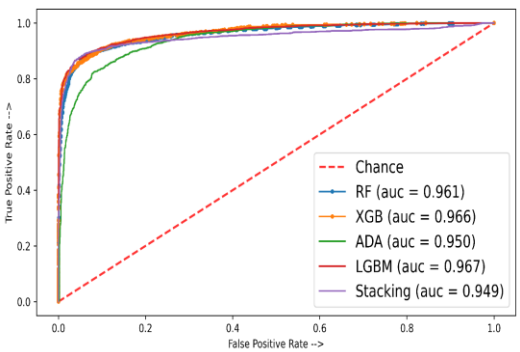

FIGURE 16. **ROC for 10 fold using CTGAN Generated Data.**

TABLE XI. exhibits the results obtained from 10-fold Cross-validation by using SMOTE architecture.

TABLE XI
RESULTS FOR SMOTE GENERATED DATA USING 10-FOLD CV

Classifier	Accuracy	Precision	Recall	**MCC**	F1	ROC
LGBM	92.26	92.61	92.32	85.18	92.38	0.979
XGB	92.47	92.48	92.45	84.96	92.41	0.979
RF	93.38	93.78	93.21	**87.57**	93.27	0.981
ADA	88.80	88.81	88.79	77.62	88.78	0.965
HSLR	93.41	93.22	93.27	86.86	93.34	0.979

The bagging approach RF has shown best results with MCC score of 87.57.
Figure 17. exhibits the ROC attained from 10-Fold cross validation for SMOTE architecture, where RF has outclassed remaining approaches with score of 0.981.

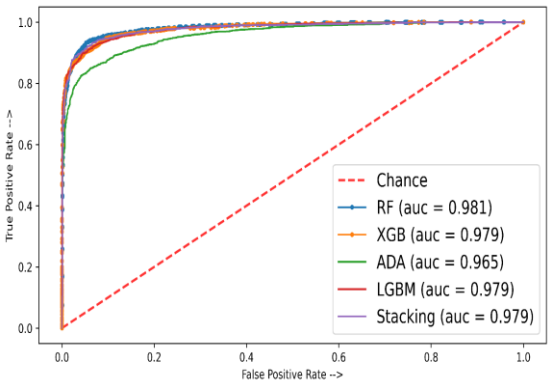

FIGURE 17. **ROC for SMOTE generated data using 10-Fold CV.**

Figure 18. demonstrates the confusion matrix by using SMOTE approach for 10-Fold cross validation. The confusion has been shown for highest results for using RF classifier.

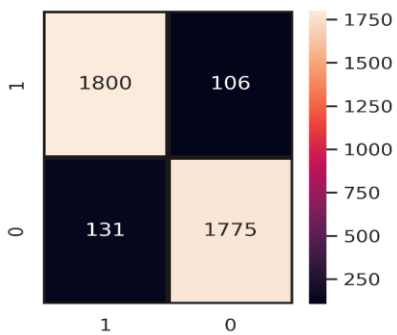

FIGURE 18. **CM from for SMOTE generated data for 10-Fold.**

V. DISCUSSION

In this section, a comparative analysis is presented between the proposed solution and existing state-of-the-art studies. The emphasis is on highlighting the enhanced performance of the proposed methodology across various metrics. Table XII offers a detailed comparison with recent studies, showcasing the effectiveness of the proposed predictor, which surpasses other methodologies. Notably, the table highlights the improved performance metrics achieved using data generated by SMOTE compared to data generated by GAN. The initial dataset had a significant class imbalance, with an unequal distribution of positive and negative classes. Such an imbalance can lead machine learning or deep learning models to be inherently biased towards the majority class, often skewing predictive accuracy.

To address this, two cutting-edge class-balancing techniques were employed: SMOTE and GAN, with a specific focus on the CTGAN architecture for synthetic data generation. This strategy not only balanced the dataset, improving the reliability of the predictive model but also provided a deeper insight into the data patterns

TABLE XII
COMPARISON WITH STATE-OF-THE ART STUDIES

Study	Accuracy	Precision	Recall	F1-Score	MCC	ROC
[16]	93.15	-	-	-	-	-
[17]	91.00	-	-	-	-	-
[19]	86.0			86.0		
Proposed Method	94.06	94.23	94.28	94.05	88.13	0.984

Our analysis revealed a discernibly better performance with The analysis indicates that SMOTE outperforms CTGAN in addressing the dataset imbalance. While CTGAN is adept at generating synthetic data, it sometimes struggles to capture complex patterns and relationships present in real data, particularly in cases of significant data imbalance and high-dimensional datasets. In contrast, SMOTE creates data points that mirror existing entries, providing a truer representation of the actual data. Its straightforward application across diverse datasets makes it the preferred choice for this study.

The proposed hybrid model, HSLR, incorporates machine-learning classifiers such as RF, XGB, ADA Boost, and LGBM as base learners, with LR acting as a meta-classifier. This combination of algorithms capitalizes on the strengths of each classifier, resulting in a model with superior predictive accuracy and dependability. In summary, this study presents a significant advancement in addressing imbalanced datasets, demonstrating a predictive model that excels in comparison to existing state-of-the-art studies. The achieved performance metrics validate the effectiveness of the approach, suggesting promising avenues for future research in this area.

The findings of this study, while promising, open up several avenues for future research and exploration in the realm of customer personality analysis and churn prediction. One potential area of exploration is the integration of deep learning algorithms, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These algorithms can be particularly effective in handling unstructured data and might offer enhanced predictive accuracy. As the field of customer personality analysis evolves, there's scope to explore additional features and attributes that might influence churn prediction. Advanced feature engineering techniques can be employed to extract more meaningful insights from the data.

With the exponential growth of data in today's digital age, ensuring that the proposed methodologies are scalable becomes paramount. Future research can focus on optimizing the current architecture to handle vast datasets efficiently, possibly integrating distributed computing frameworks like Apache Spark. One of the potential areas of exploration is the development of real-time churn prediction systems, providing businesses with immediate insights and allowing them to take proactive measures to retain customers. The current study, while focused on a specific industry or domain, leaves room for exploration of the applicability of the proposed methodologies across different industries, understanding the nuances and challenges unique to each.

As synthetic data generation techniques become more prevalent, addressing ethical considerations related to data privacy and usage will be crucial. Future research can delve into developing frameworks that ensure the ethical generation and use of synthetic data. The integration of traditional statistical methods with machine learning and deep learning algorithms can lead to the development of hybrid models, offering a more holistic view of customer behavior. Based on the insights derived from customer personality analysis, future research can also focus on devising personalized marketing strategies tailored to individual customer preferences, enhancing engagement and loyalty. Incorporating a feedback loop mechanism can ensure that the models are continuously updated based on real-world performance, leading to more adaptive and resilient prediction systems. In conclusion, the field is ripe for further exploration, with research in this domain playing a pivotal role in shaping customer-centric strategies and ensuring sustained growth.

VI. CONCLUSION

This study tackled the challenge of class imbalance in machine learning models by employing CTGAN and SMOTE to generate synthetic data. The results indicated SMOTE's superiority over CTGAN in terms of various performance metrics. The dataset used exhibited class imbalances, which can bias machine learning models towards the majority class. This issue was addressed by generating synthetic data using SMOTE and CTGAN. The proposed HSLR model utilized various classifiers, and its performance was evaluated using metrics such as accuracy score, precision, recall, F1 score, MCC, and ROC score. The SMOTE approach yielded the highest results, outperforming existing methods. Future plans include collecting more data and exploring deep neural architectures like FCN, CNN, LSTM, and GRU. This study's findings offer insights into the application of machine learning and deep learning in addressing class imbalance issues, with potential applications in domains like healthcare, finance, and security. The code is available on the GitHub repository: https://github.com/mazhar786/Customer-personality-.

The notational Table XIII of each abbreviation used is as below

TABLE XIII
NOTATIONAL TABLE

NOTATION	DESCRIPTION
RF	RANDOM FOREST CLASSIFIER
XGB	EXTREME GRADIENT BOOSTING
ADA B	ADABOOST
LGBM	LIGHT GRADIENT BOOSTING MACHINE
LR	LOGISTIC REGRESSION
SMOTE	SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE
GAN	GENERATIVE ADVERSARIAL NETWORK
CTGAN	CONDITIONAL TABULAR GENERATIVE ADVERSARIAL NETWORK
CNN	CONVOLUTIONAL NEURAL NETWORK
GRU	GRADIENT RECURRENT UNIT

REFERENCES

[1] L. Marin, S. Ruiz, and A. Rubio, "The role of identity salience in the effects of corporate social responsibility on consumer behavior," J. Bus. Ethics, vol. 84, no. 1, pp. 65-78, 2009.
[2] E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A neural network ensemble with feature engineering for improved credit card fraud detection," IEEE Access, vol. 10, pp. 16400-16407, 2022.
[3] S. A. Ebiaredoh-Mienye, E. Esenogho, and T. G. Swart, "Artificial neural network technique for improving prediction of credit card default: A stacked sparse autoencoder approach," Int. J. Electr. Comput. Eng., vol. 11, no. 5, pp. 4392, 2021.
[4] M. J. Awan, M. A. Khan, Z. K. Ansari, A. Yasin, and H. M. F. Shehzad, "Fake profile recognition using big data analytics in social media platforms," Int. J. Comput. Appl. Technol., vol. 68, no. 3, pp. 215-222, 2022.
[5] C. Leuz, "Evidence-based policymaking: promise, challenges and opportunities for accounting and financial markets research," Account. Bus. Res., vol. 48, no. 5, pp. 582-608, 2018.
[6] M. J. Awan, M. S. M. Rahim, H. Nobanee, A. Munawar, A. Yasin, and A. M. Zain, "Social media and stock market prediction: a big data approach," Comput. Mater. Continua, vol. 67, no. 2, pp. 2569-2583, 2021.
[7] K. Chaudhary, M. Alam, M. S. Al-Rakhami, and A. Gumaei, "Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics," J. Big Data, vol. 8, no. 1, 2021.
[8] A. S. Kumar and D. Chandrakala, "A survey on customer churn prediction using machine learning techniques," Int. J. Comput. Appl., vol. 975, p. 8887, 2016.
[9] N. N. Y. Vo, S. Liu, X. Li, and G. Xu, "Leveraging unstructured call log data for customer churn prediction," Knowl.-Based Syst., vol. 212, p. 106586, 2021.
[10] J. Sun, Z. Tian, Y. Fu, J. Geng, and C. Liu, "Digital twins in human understanding: a deep learning-based method to recognize personality traits," Int. J. Comput. Integr. Manuf., vol. 34, no. 7–8, pp. 860–873, 2021.
[11] N. Chaudhuri, G. Gupta, V. Vamsi, and I. Bose, "On the platform but will they buy? Predicting customers' purchase behavior using deep learning," Decis. Support Syst., vol. 149, p. 113622, 2021.
[12] H. Zhao, Z. Liu, X. Yao, and Q. Yang, "A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach," Inf. Process. Manag., vol. 58, no. 5, p. 102656, 2021.
[13] E. Utami, I. Oyong, S. Raharjo, A. D. Hartanto, and S. Adi, "Supervised learning and resampling techniques on DISC personality classification using Twitter information in Bahasa Indonesia," Appl. Comput. Inform., 2021.
[14] M. Hassanein, W. Hussein, S. Rady, and T. F. Gharib, "Predicting personality traits from social media using text semantics," in 2018 13th Int. Conf. Comput. Eng. Syst. (ICCES), 2018, pp. 184–189.
[15] A. A. Tudoran, "A machine learning approach to identifying decision-making styles for managing customer relationships," Electron. Mark., pp. 1–24, 2022.
[16] R. Hegde, S. K. Hegde, S. Kotian, and S. C. Shetty, "Personality classification using data mining approach," IJRAR19H1202 Int. J. Res. Anal. Rev., vol. 354, no. 1, pp. 354–359, 2019.
[17] A. Sharma, A. Pratap, K. Vyas, and S. Mishra, "Machine learning approach: Consumer buying behavior analysis," in 2022 IEEE Pune Sect. Int. Conf. (PuneCon), 2022, pp. 1–10.
[18] S. De, P. P, and J. Paulose, "Effective ML Techniques to Predict Customer Churn," in Proc. IEEE ICIRCA, Sep. 2021.
[19] S. C. Koumetio Tekouabou, Ş. C. Gherghina, H. Toulni, P. Mata, and J. Martins, "Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods," Mathematics, vol. 10, no. 14, pp. 2379, Jul. 2022.
[20] A. Chowdhury, S. Kaisar, M. Rashid, S. S. Shafin, and J. Kamruzzaman, "Churn Prediction in Telecom Industry using Machine Learning Ensembles with Class Balancing," in Proc. IEEE CSDE, Dec. 2021.
[21] "Customer Personality Analysis," Kaggle: Your Machine Learning and Data Science Community. [Online]. Available: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis. Accessed: Apr. 10, 2022.
[22] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," Adv. Neural Inf. Process. Syst., vol. 32, 2019.
[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
[24] L. Breiman, "Random forests," Mach. Learn., vol. 45, pp. 5–32, 2001.
[25] Y. Freund, R. E. Schapire, and others, "Experiments with a new boosting algorithm," in icml, 1996, pp. 148–156.
[26] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu, and W. Zeng, "Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data," Agric. Water Manag., vol. 225, pp. 105758, 2019.
[27] L. Torlay, M. Perrone-Bertolotti, E. Thomas, and M. Baciu, "Machine learning–XGBoost analysis of language networks to classify patients with epilepsy," Brain Inform., vol. 4, no. 3, pp. 159–169, 2017.
[28] I. Düntsch and G. Gediga, "Confusion Matrices and Rough Set Data Analysis," J. Phys. Conf. Ser., vol. 1229, no. 1, 2019.21