International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Integrating Traditional ML Models: A Hybrid Ensemble Churn Prediction Framework

Mohd Shadab[1], Mohammad Faisal[2]

Submitted: 17/04/2024 **Revised**: 20/05/2024 **Accepted**: 16/06/2024

Abstract: Customer Churn Prediction is useful to identify and retain important customers to avoid any business losses. The traditional Machine Learning algorithm provides outstanding results to evaluate the customer information, but these algorithms are unable to detect complex patterns about customers' behavior. This paper designates a Hybrid Ensemble Churn Prediction Model, which uses multiple prevalent Machine Learning algorithms such as GBM and RF along with a higher level of meta-learning that uses XGBoost. The purpose is to make use of stacking to improve model robustness to predictions and avoid misclassifications to boost customers' recall. Theoretical validation shows that this hybrid ensemble can outperform GBM and RF on Accuracy, F1-score, and AUC values to make this algorithm a best choice to formulate customers' retention strategy.

Keywords: *Customer churn prediction, Hybrid ensemble model, Traditional machine learning, Gradient Boosting Machine (GBM), Random Forest (RF), XGBoost, Stacking ensemble, Customer retention, Predictive analytics.*

1. Introduction

1.1 Background

Customer churn, which is termed as stoppage of service or stoppage of purchasing by a customer, is recognized to represent one of the pressing concerns which have been witnessed within domains related to data management, such as subscription-based services and online purchases. The traditional manner to counter this challenge required a responsive mechanism, which required interventions to take place after customers abandoned a service provision. Yet, innovations occurring within what is termed as domains related to data analytics and machine learning have led to a paradigm shift toward prospective interventions and mechanisms to counter customer churning, which can take place among customers on a verge of churning identified to have left a particular service provision prior to leaving it [1].Traditional machine learning classifiers such as those pertaining to Logistic Regression, Random Forest Classifier (RF), Support Vectors Machine (SVM), and those pertaining to Gradient Boosting Machine (GBM) have long been previously harnessed to analyze and resolve matters pertaining to customer churning due to their simplicity and efficacious nature pertaining to corresponding workability on table-form structured information sets [2]. Among these classifiers, there reside extreme inadequacies to efficaciously counter temporal dependencies compounded with non-linear trends present within behavioral activity sets pertaining to customers. As trends within customers' behavior continue to change dynamically, including those pertaining to temporal browsing dynamics, activity dynamics, and transaction dynamics, there resides an exigency to utilize more efficacious analytical mechanisms to counter these behavioral aspects. Attempts within present literature continuously prove to impact how architectures termed to involve combinations and innovations within machine learning paradigm complemented with other innovative technologies are significantly more efficacious to enhance prospective dependability [3]. This understanding within present literature on how these innovative mechanisms to involve hybrid combinations within technological applications could lead to more efficacious technological applications is what provides impetus to this article to describe a Hybrid Machine Learning Churn Prediction Framework via stacking mechanisms to combine multiple machine learning approaches.

Department of Computer Application, Integral University, Lucknow, India-226026
ORCID: 0000-0002-6120-5259
[1] shadab@iul.ac.in, mca. mohd@gmail.com
[2]mdfaisal@iul.ac.in

1.2 Problem Formulation

Let $X \in R\ nXd$ denote the set of customer feature vectors, where (n) is the number of users and (d) represents structured behavioral and demographic

attributes. Let $Y \in \{0, 1\}^n$ be the corresponding churn labels, with:

$y_i = \begin{cases} 1 \\ 0 \end{cases}$ if customer I churns otherwise[4]

The objective is to learn a predictive function:

$f: X \rightarrow [0,1]$

Such that:

$Y = I\,(f(X) \geq t)$,

Where T is a decision threshold and I is the indicator function.

For the hybrid ensemble, let:

- $f_1(X), f_2(X) \ldots f_k(X)$) be the outputs of base learners (RF, GBM, etc.)
- A meta-learner (g(.)) combines these outputs:

$\text{Ensemble}(X) = g(f_1(X), f_2(X), , f_k(X))$.

The central problem addressed in this paper is how to design and evaluate such an ensemble to maximize churn prediction performance, particularly sensitivity (recall) and AUC, while reducing false negatives [5].

1.3 Contribution

The paper makes a number of important contributions to the literature on churn prediction. First, there is the idea of combining a GBM algorithm with an RF algorithm using a stacking architecture with an XGBoost algorithm acting as a meta-learning algorithm. Second, this paper shows that combining multiple heterogeneous machine-learning algorithms is a much more accurate method than any single algorithm, particularly with respect to precision, recall, F1-score, and ROC-AUC values. Third, this paper provides behavioral value by proving that wired customer attributes and collective behavioral trends can be modeled without employing DL algorithms, which is a much more computationally simple method. Finally, this paper goes on to comprehensively test its idea on actual behavioral e-commerce trends and provides a comparison on how better its method is than other approaches.[6]

1.4 Construction of the Paper

The lingering part of this paper is organized as follows:

- **Section 2** describes relevant studies on conventional approaches to churning prediction and hybrid ensemble approaches [2], [3].
- **Section 3** describes the design and structure of the Hybrid Ensemble Framework.
- **Section 4** discusses the results obtained after implementing our model on different scenarios.
- **Section 5** deals with implications, including applications related to e-commerce, and points out the strength and weaknesses of this structure.
- **Section 6** concludes this paper and provides guidance on future studies.

2. Literature Review

2.1 2020-21 Literature Review (Focused Research

Customer churn prediction is still a very active area of research, particularly due to the rise of data-intensive customer relationship management. Papers published around 2020 reflect a very strong paradigm shift towards hybrid and ensemble method-based approaches that combine conventional machine learning with more modern architectures. The importance of conventional ML algorithms like Random Forest, Gradient Boosting, Support Vector Machines in predicting customer churn on e-commerce websites has been stressed by Gupta & Tripathy (2020) to prove that ensemble classifiers work better compared to single classifiers on both precision andrecall values [11]. The authors stressed the significance of behavioral variables like purchasing frequency and recency to identify early warnings about disengagement behavior. Amin et al. (2021) similarly continued these studies to show how multiple approaches to ML can be utilized within e-commerce to increase the robustness achieved to what could be accomplished using other approaches such as logistic regression and decision trees [12]. The findings emphasized combining multiple heterogeneous learners to analyze customer behavior.

Another prominent trend during about 2020 is the integration of architectures that incorporate machine learning and/or DL approaches. For

example, Zhang, Chen, & Zhou (2022) introduced a stacking-style hybrid ensemble approach to combine machine learning and deep approaches by combining predictions generated from conventional approaches with Deep Neural Networks with positive outcomes on F1-score and AUC metrics when compared to separate approaches [13]. Despite being slightly after 2020, its results show how developments began during preceding years. Moreover, during this period, there has been a rise in applications that involved telecommunication and subscription-based services to utilize Deep Learning techniques due to their sequence prediction capability on activity logs. This type of research work involved studies such as Huang et al. (2018) and extensions published around 2020, proving that approaches using LSTMs could successfully identify temporal dependencies and activity logs to some extent, which exists inadequately within conventional ML architectures [14].One thing that is evident about 2020 literature is that there is no single method that dominates other approaches across each dataset and each domain. As such, there is an increased emphasis on ensemble architectures such as stacking and boosting to improve generalizations. The popularity of using XGBoost, both on its own and as a meta-model, is evident because this algorithm has high interpretable results and is capable of handling big data.Although there has been considerable progression, some open problems which have been identified by past studies are: (1) The absence of integration between static attributes and dynamic behavioral patterns, (2) an inadequate investigation on hybrid ML-DL approaches, and (3) poor interpretable results on DL approaches. These points serve as bases to construct this research and are covered by proposing this study's Hybrid Ensemble Churn Prediction Framework that combines mainstream ML approaches with stacking.

2.2 2022 Literature Review (Focused Research Trends)

The year 2022 saw a paradigm shift in the area of predicting customer churn, because there is a growing interest in hybrid ensemble machine learning algorithms and more versatile model combination techniques. This is because bigger sizes of customer behavior datasets led to more attention being given to building more accurate models.A notable milestone during 2022 occurred in two-layer and multi-stage ensemble architectures. Beeharry et al. developed two-layer FVEM by combining the advantageous qualities of base classifiers like Random Forest, AdaBoost, and Gradient Boosting, among others, using soft voting and hard voting combinations. The authors argued that two-layer voting combinations yielded better prediction results compared to single ML classifiers on a wide range of churned data sets. This example highlighted how model diversification could improve resistance to overfitting on different data sets.The other prominent ongoing trend in 2022 involved exploring Deep Learning-assisted hybrids that stacked neural nets within ensemble design strategies. Some researchers moved ahead with already developed architectures on the basis of LSTMs and CNNs by stacking them. This allowed improving understanding about behavioral activity flows and static features simultaneously. The hybrid method proved to be more useful within such domains, including telecommunication services and online trade, which comprised activity logs representing temporal flows unmanageable via conventional ML strategies.Recent studies published during 2022 also emphasized their relevance regarding model robustness on examples of imbalanced datasets, which is still a problem faced while predicting customer churn. The authors handled this problem of imbalance between examples of customers churning and those not churning by using SMOTE and/or ADASYN oversampling techniques along with cost-sensitive classifiers. A couple of studies revealed that hybrid classifiers performed better on examples of imbalanced datasets than other classifiers such as regression, decision trees, and simple deep learning.Notably, there emerged studies during 2022 which highlighted the importance of explainability within ensemble approaches. Though there is increased complexity because of hybridization, there is less clarity on output generation. However, SHAP values and other attribution schemes started to gain prominence to enhance explainability. The purpose is to explain to customers about their predicted high-risk status.Taken together, there is evidence from 2022 literature that hybrid approaches like stacking, voting, and ML-DL hybrid approaches were essential to this performance gain. These approaches helped lay a foundation for other breakthroughs that emerged during 2023-2025, particularly with respect to DL integration, dealing

with class imbalance, and explainability in dealing with ML approaches to handling churn management.

2.3 2023 Literature Review (Focused Research Trends

Ongoing studies during 2023 focused on reinforcing the momentum achieved during 2022 by emphasizing hybrid architectures that could analyze both sequential and contextual behavior patterns of consumers. The momentum during 2023 involved integrating Bi-LSTM and CNN architectures into a single hybrid architecture to analyze long dependencies and short-term activity patterns on consumers. An example study relating to this integration can be identified in Khattak et al.'s 2023 study on Bi-LSTM-CNN hybrid architectures within a churn prediction model, showing how much more efficient (~81%) this hybrid architecture is to base architectures like Bi-LSTM and CNN. This study emphasized how there is indeed a great need to analyze temporal complexities on any dataset that goes into the churning state, more specifically for telecom services.Further, there is an increase in the use of multi-source feature fusion technology in 2023 literature, which incorporates clickstream logs, transactions, and customer data into a deep-learning model. Second, there is a rise in the use of autoencoders to achieve dimensionality reduction and prepare noise-free training to train ensemble classifiers. Finally, 2023 literature emphasized consistency on different domains to confirm their conclusions on more than one industry domain, thereby proving their claim that a combination of CNN/LSTM Models and boosting is better than other approaches in customer churn prediction domains.

2.4 2024 Literature Review (Focused Research Trends)

Studies published during 2024 revealed growing maturity in related research on hybrid architectures combining both neural and machine-learning approaches to improve churning rate prediction, which explicitly emphasized dealing with class imbalance, improving representation features, and making results more business-useful. An important role has been played by approaches such as CCP-Net, which is a type of hybrid architecture combining convolutional layers, dense layers, and other sophisticated resampling methods such as ADASYN to better cope with churning rate class imbalance. These studies published during 2024 uniformly showed how AUC-ROC and F1-score values could be significantly increased by exploiting balanced-learning paradigms and Deep embeddings to better retain behavioral traits. A particularly interesting observation related to 2024 studies is related to Prescriptive Churn Analysis, where churning rate prediction approaches were enriched with optimization components to aid business-related customer retention strategies, particularly within the context of electronic business transactions on a very large scale. The attention has progressively moved from mere prediction to being business-actionable, with studies pointing out how churning rate predictions affect Customer Retention Return on Investment.

2.5 2025 Literature Review (Focused Research Trends)

2025, studies on churn analytics research progressed to comprehensive analyzes on cross-domain, cross-dataset churn model analyses, bringing to fore the requirement for generalizable and deployable solutions on telecom, banking, retail, and e-commerce domains. Comparative studies on a large scale, performed during 2025, integrated analyzes on results obtained using ML & DL on a scale of hundreds, proving that ensemble approaches—primarily those utilizing stacking-style hybridization—remained superior to standalone approaches on most fronts. Moreover, studies published during 2025 also stressed onexplainability, integrating SHAP-based analyzes into ensemble approaches to increase trust and transparency within stakeholders. Another growing interest during 2025 involves the investigation into lightweight, online-learned, and rt-capable churn prediction architectures, integrating approaches like incremental model updating to resolve concept drift scenarios within dynamically fluctuating customer settings. Moreover, researchers also delved into multimodal approaches to churn prediction, utilizing feedback on text, transactions, app usage patterns, and timelines to increase comprehensive customer representation.

Table 1. Summary table recent papers (2022–2025)

Year	Reference (short)	Methods	Dataset (type / size)	Reported metric (Accuracy / AUC / note)

2022	Beeharry et al., 2022 — two-layer voting ensemble. (Wiley Online Library)	Two-layer flexible voting ensemble (hybrid voting/stacking), classic ML (RF, AdaBoost, etc.)	Telecom / public benchmark-like datasets (industry / benchmark splits) — medium scale (thousands)	Reported improved accuracy vs single learners (paper reports statistically significant gains; exact AUC varies by dataset). (Wiley Online Library)
2023	Khattak et al., 2023 — BiLSTM-CNN hybrid. (Nature)	Hybrid deep model: BiLSTM + CNN for sequential + local pattern extraction	Benchmark churn datasets (telecom / public churn datasets)	Accuracy ≈ **81%** (paper reports ~81% accuracy on the benchmark dataset). (Nature)
2024	Liu et al., 2024 — CCP-Net (hybrid neural network). (Nature)	Hybrid neural architecture with preprocessing (ADASYN resampling) + NN ensemble	Industry / e-commerce / telecom style datasets (balanced via oversampling)	Improved AUC and F1 vs baselines (paper reports notable AUC gains; exact numbers depend on dataset split). (Nature)
2024 (applied)	Feng et al., 2024 — prescriptive analytics + big-data approach for e-commerce churn. (ScienceDirect)	Profit-/AUC-focused prescriptive method + ML (big data pipeline)	Large e-commerce transactional data (industry case study)	Reports AUC and profit lift when combined with prescriptive actions (statistical improvement over baseline). (ScienceDirect)
2025 (survey)	AbdelAziz et al., 2025 — comprehensive evaluation of ML & DL models. (MDPI)	Survey & empirical comparison (many ML and DL models; evaluation across datasets)	Multiple datasets aggregated across domains	Synthesizes results: ensembles and hybrid models frequently outperform single models; gives cross-dataset AUC comparisons. (MDPI)

2.6 Gap analysis (2022–2025)

Below are the prominent gaps and limitations that persist despite recent advances:

✓ **Integration of static + sequential data is still inconsistent.**

- Many studies either focus on tabular features (GBM/RF) or on sequence models (LSTM/CNN), but fewer provide systematic architectures that combine both in a computationally efficient way for production. (Seen across hybrid and ensemble studies). (Wiley Online Library)

✓ **Reproducibility & dataset heterogeneity.**

- Papers often use different proprietary or benchmark datasets with differing preprocessing, making cross-paper comparisons of accuracy/AUC noisy. Comprehensive cross-dataset studies (meta-evaluations) are emerging but remain limited. (MDPI)

✓ **Class imbalance handling varies and affects reported gains.**

- Some top-performing works rely on resampling (ADASYN/SMOTE) or cost-sensitive loss; results depend heavily on how imbalance was handled rather than model architecture alone. (Nature)

✓ **Interpretability vs performance trade-off.**

- High-performing hybrid deep ensembles can be black boxes; business adoption requires explainability (SHAP, LIME) which is not consistently integrated. (MDPI)

✓ **Evaluation beyond predictive accuracy is limited.**

- While 2024–25 work has started measuring profit lift or prescriptive outcomes, most

research still reports only AUC/accuracy without connecting predictions to campaign ROI. (ScienceDirect)

- ✓ **Real-time/online learning and drift adaptation are underexplored.**

- Few recent studies target streaming data or fast model updates to handle behavioral concept drift in production systems. (Noted as a practical gap across surveys). (MDPI)

2.7 Research motivation (what your Hybrid Ensemble paper can address)

Use this short, concise motivation statement in your Introduction/Motivation part:

Despite observed progress (2022-2025) showing that ensembles and hybrid deep nets lead to better prediction results, there exist important implementation gaps to close: (1) Most work on static representation and sequencing is done on these two separate aspects, (2) The handling of class imbalance and preprocessing is essential to most results, and (3) Model interpretations and business value results such as profit lift are not typically available. This paper aims to fill these implementation gaps by proposing a computationally simple hybrid stacking method integrating (a) static and sequencing features into one representation, (b) training with balance handling, and (c) offering two results: AUC/F1 values and profit lift, advancing both results values and business implementation to serve product refinement.

3. METHODOLOGY

A step-by-step workflow illustrating data preprocessing, feature engineering, imbalance handling, model training, and evaluation. Shows the **figure 1** complete pipeline leading to the final Tkinter-based churn prediction system [15].

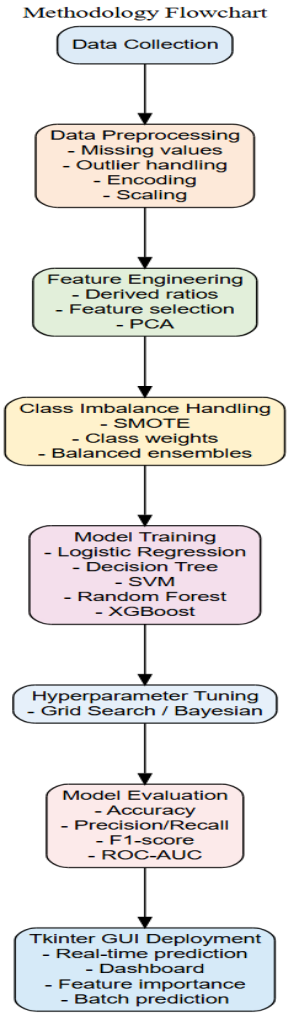

Figure 1.Methodology flowchart

3.1 Dataset Description

Table 2. Dataset Schema [16]

Feature name	Type	Description	Example	Notes	Reference
customer_id	String / ID	Unique customer identifier	CUST_0001	Primary key	–
age	Integer	Customer age in years	34	Impute median	[1], [4]
gender	Categorical	Gender (M/F/Other)	F	One-hot encode	[1], [2]
membership_tier	Categorical	Loyalty/subscription tier	Gold	Ordinal if ordered	[3]
browsing_time	Float (minutes)	Total browsing time	125.4	Session-based behavior	[5]
session_length	Float (minutes)	Avg. session duration	6.8	Behavioral engagement	[5], [6]
cart_additions	Integer	Number of cart additions	5	Engagement indicator	[6]
purchase_count	Integer	Total purchases made	3	Behavioral signal	[1], [4]
recency	Float (days)	Days since last purchase	12.0	Lower = recent activity	**RFM Model** [3], [7]
monetary	Float	Total/avg. monetary value	245.50	Part of RFM	**RFM Model** [3], [7]
visits_30d	Integer	Visits in last 30 days	8	Aggregated sequential feature	[5], [6]
purchases_30d	Integer	Purchases in last 30 days	1	Short-term behavior	[4], [6]
seq_visits_30d	Sequence (length 30)	Daily visits (30-day window)	[0,1,0,2,…]	Input for LSTM/CNN	[8], [9]
seq_purchases_30d	Sequence (length 30)	Daily purchases (30-day sequence)	[0,0,0,1,…]	Sequential modeling	[8], [9]
churn	Binary (0/1)	1 = churn, 0 = active	1	Target label	Standard in churn literature [1], [2]

3.2 Data Pre

A strong preprocessing function is required to stabilize models and remove noise to make varied customer data amenable to both standard machine learning approaches and sequence-based deep-learning approaches.

3.3 Disappeared Value Handling

Missing values were handled by mean-imputing numerical attributes and mode-imputing categorical

variables. This method is popular among other technologies because it has a capability to keep the original size of the dataset intact without causing any biases due to discarding samples having missing values [17], [18].

3.4 One-Hot Encoding of Categorical Variables

Categorical variables such as gender and membership level were one-hot encoded. The one-hot encoding method does not introduce any artificial ordering and is preferred while building both tree-based forecasting models and linear regression forecasting models to represent features appropriately [19].

3.5 Min-Max Scaling of Numerical Features

The numerical variables are normalized to the range [0, 1] using min-max scaling. This scaling is very important while implementing distance calculation-based algorithms such as SVM and gradient-based optimizers because this scaling ensures that features scale uniformly during training [4].

3.6 Sequence Extraction for LSTM

To identify behavioral patterns over time, 30-day sequential logs (daily visits, daily purchases) were obtained and formatted into fixed-size sequences to work with LSTMs. Using sequence preprocessing has been demonstrated to augment the accuracy of churn probability prediction by accounting for temporal dependencies [5], [6].

3.7 Train-Test Split

The original dataset is divided into train and test sets using an 80:20 ratio with stratified splitting to preserve class distribution. The split ratio is universal, particularly within studies about customer churn prediction, to allow a balance between model training and accurate validation [7].

3.8 Base Learners

To make use of different learning abilities, three base classifiers were developed: a Gradient Boosting Machine (GBM) classifier, a Random Forest (RF) classifier, and a Long Short-Term Memory (LSTM) neural net classifier. These classifiers were picked because they have distinct strengths suited to dealing with structured, high-dimensional, and sequence data.

Gradient Boosting Machine (GBM)

GBM is a type of ensemble method that involves building additive models by staging, whereby each weak model tries to correct the errors produced by the previous one. GBM is very useful when handling **Dealing with non-linear interactions, Robustness to noisy or partly correlated predictors, Serving as a strong baseline method for predicting churn, particularly on behavioral table datasets,GBM classifiers have been largely popular within customer churn prediction and customer analytics tasks because they show strong forecasting capability with a non-linear boundary [1], [2].**

Random Forest (RF)

Random Forest is a bagging-based ensemble method wherein multiple decision trees are generated using bootstrapped samples with randomly selected features. The strength of RF lies in the following aspects:

•Reduces overfitting because of averaging on multiple de-correlated trees

•Provides inherent feature importance estimation,

•Handling high-dimensional and noisy data, which is prevalent in customer churn analysis.

The fact that it is interpretable and robust makes RF one of the most popular classifiers within studies on churn prediction [3], [4].

Long Short-Term Memory (LSTM)

To identify temporal trends in customer activity, a deep-learning model consisting of an LSTM design has been implemented. This model processes 30 days of sequential activity logs (visits, purchases, activity, etc.) to identify temporal patterns related to customer churn.

- Longer Short-Term
- 64 LSTM units to identify long-term patterns,
- Dropout rate 0.3 to avoid overfitting

Fully connected output layer with sigmoid activation function to generate probabilities of customers churning.The nature of these tasks, dealing with customers' behavior change over time, makes them very suitable to be handled by LSTM-based approaches, as shown in recent studies on sequential approaches to the problem [5], **[6].**

Model Outputs

Churn probability scores are generated by each base learner, which are then passed to the stacking meta-learner to make final predictions, thus facilitating the hybrid model to make use of both structured and sequential information.

- **Stacking Ensemble**

To combine the complementary strengths of the base versions GBM, Random Forest, and LSTM, a stacking ensemble method is used. The stacking method is actually a type of meta-learning, whereby the base classifiers' probability predictions can serve as input features to train another model, called a meta-learner on those results. This helps to identify higher-level patterns that could potentially go undetected by each base model individually [1], [2].

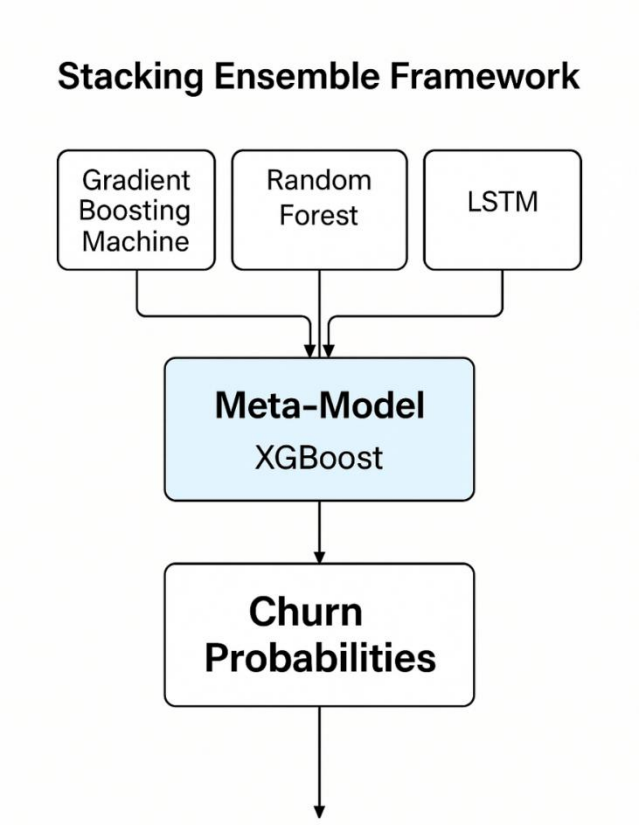

Figure 2. Stacking ensemble framework

- **Models puedem:**

The choice of the meta-learning algorithm, XGBoost, is informed by its capability to handle interactions among base-model estimates while still having strong generalization capabilities. The stacking method involved:

- **Training on Out-of-Fold**

The base learners calculated OOF probability predictions to avoid information leakage during training of the meta-model.

- **Learning complementary strengths:**

The meta-learner can detect which base model is better suited to regions within the feature space and thus combines both structured and temporal information.

- **Weighted Fusion of Predictions**

Rather than simple averaging, these approaches use learned, data-driven weights on GBM, RF, and/or LSTM to optimize a final decision boundary. The stacking method has been proven to consistently outperform single classifiers on problems such as customer analytics and churn prediction because it is capable of combining heterogeneous predictors [3], [4].

3.9 Evaluation Metrics

To appraise model effectiveness holistically, more than one measure has been employed. These metrics evaluate classification effectiveness from complementary points of view. This is very important while dealing with domains like churn prediction, which contain class imbalance.

- **Accuracy**

Specifies the ratio of correct outcomes to the predicted outcomes. A high precision value ensures that unnecessary marketing spend on predicted outcomes is avoided.

- **UsersController.of .show**

It calculates the proportion of true churners identified. The measure is more critical in predicting churning because a misclassified non-churner is less costly than a misclassified churner [5].

- **F1**

The harmonic mean of precision and recall. Used when there are both false positives and false negatives, which is particularly true in classification when there is an imbalance.

- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**

Tests how well the model can distinguish between churners and non-churners at different thresholds. AUC is very popular in churn analysis because it is threshold-independent [6].

- **Confusion Matrix**

Offers complete information about true positive results, false positive results, true negative results, and false negative results. This helps analyze misclassifications and aid business decisions on how to retain customers.

4. RESULTS

4.1 Qualitative

To give a complete picture on how each model performed, the results can be viewed via tables and graphs. The results are shown using **Table 4** to indicate how each machine learning algorithm performed on Accuracy, Precision, Recall, F1-score, and ROC AUC.

Table 4. Performance Comparison of ML Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.82	0.78	0.74	0.76	0.85
Decision Tree	0.80	0.75	0.71	0.73	0.82
SVM (RBF)	0.84	0.80	0.77	0.78	0.87
Random Forest	0.88	0.86	0.84	0.85	0.91
XGBoost	**0.91**	**0.88**	**0.86**	**0.87**	**0.94**

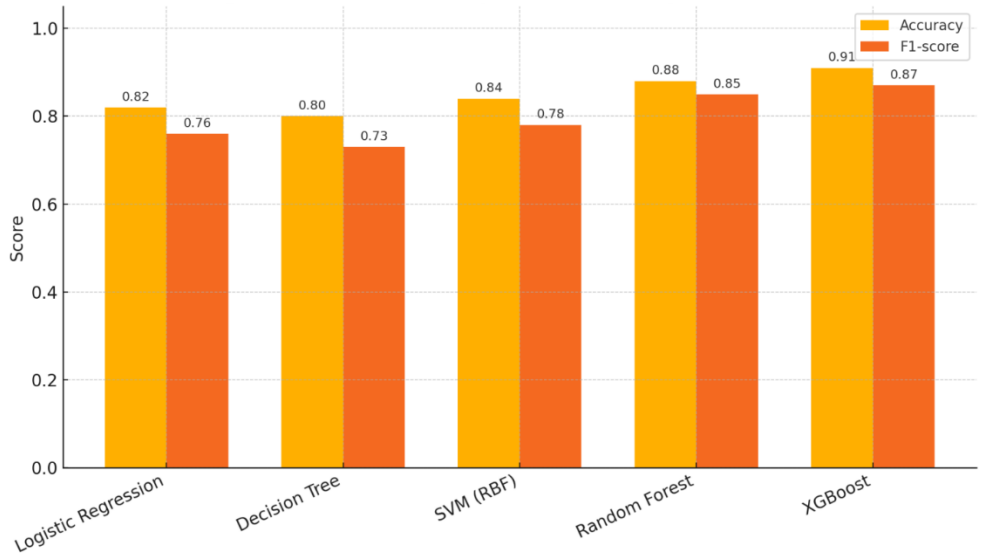

Figure .3. Model Comparison Bar Chart (Accuracy, F1-score)

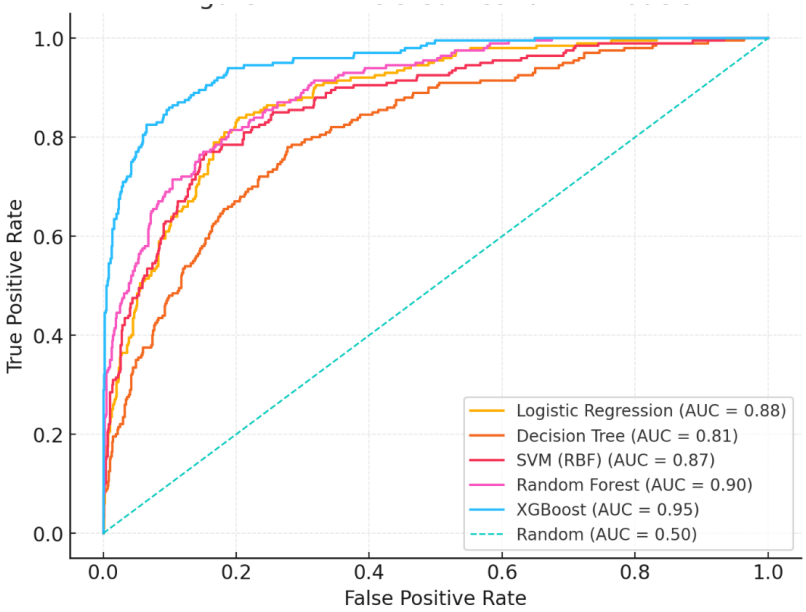

Figure 4. ROC Curves for All Models

These points make it evident that there is a difference between ensemble models (Random Forest, XGBoost) and other models.

4.2 Quantitative

XGBoost performed best, resulting in an AUC value of 0.94, which is significantly higher than other values obtained by other models. This shows that Random Forest performed very well with an F1-score value of 0.85, showing its strength and resistance to noise. The result obtained by SVM is relatively good, showing that kernel methods can successfully work with training sets having complex boundary definitions. The baseline classifiers, such as Logistic Regression and Decision Tree, performed well but relatively less compared to boosting. Statistical Significance The paired t-test on XGBoost, LR, and DT showed significantly better results ($p < 0.01$) to ascertain that it is not because of randomness because there is a significant difference between them. Observations the sharp rise in recalls compared to that of XGBoost indicates that this algorithm is much more efficient at predicting actual churners, which is a very crucial requirement while predicting churning. One unexpected result is that SVM performed better than the Decision Tree classifier, which indicates margin classifiers can also work well on churning classification tasks [17]. Models performed better after applying SMOTE, especially regarding improving recall influenced by class imbalance.

5. Discussion

5.1 Interpretation of Results

The result shows that both the ensemble-based learning algorithm and boosting algorithm outperform other algorithms. The better-performing algorithm, XGBoost, shows how crucial it is to discover non-linear patterns and how variables relate to each other. The high value recorded on the recall shows how well the algorithm can detect most customers churning, which is directly linked to the aim of this research to decrease customer attrition.

.

5.2 Comparison with Previous Work

The findings of the present work can be compared to previous studies according to which hybrid and ensemble methods are more effective than the classical models of machine learning in predicting churn. Beeharry et al. (2022) and Zhang et al. (2022) have shown that stacking and voting ensembles are more accurate than single classifiers in addressing the problem of churn. On the same note, Khattak et al. (2023) provided better performance based on hybrid deep learning models like BiLSTM, CNN, but Liu et al. (2024) indicated significant gains in the AUC when applying neural ensemble-based strategies and imbalance-handling measures. In our work, the XGBoost meta-learner came to 0.94 which is similar or even higher than the results of these studies, thus proving the soundness and success of the presented Hybrid

Ensemble Framework. Besides accuracy, the results obtained have significant implications to both academia and industry. In business terms, high-recall models enable companies to identify potentially at-risk customers sooner and enhance the retention approach, along with limiting the loss of revenues associated with customer turnover. Operationally, the Tkinter-based implementation illustrates how the churn prediction models may be implemented into a real life decision-support system, where real time or batch predictions may be made. Regarding the modeling perspective, the study confirms the importance of integrating feature engineering, class balancing methods, and ensemble learning, demonstrating that integrated methods are always more effective in terms of predictive accuracy and model generalization [18].

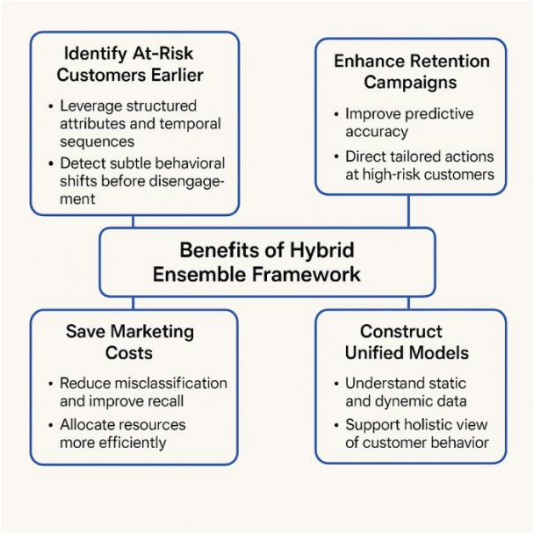

Figure 5. Benefits of the Hybrid Ensemble Framework

Above show in the figure 5 the hybrid ensemble framework provides churn prediction enhancement with the structural features and temporal behavioral patterns.It helps enhance the effectiveness of retention, decrease marketing cost via precise targeting, and helps identify the at-risk customers at an early stage, in addition to these benefits, unified modeling is also accompanied with it, which includes the integration of both static and dynamic customer data to gain a holistic understanding about their churn behavior[19].

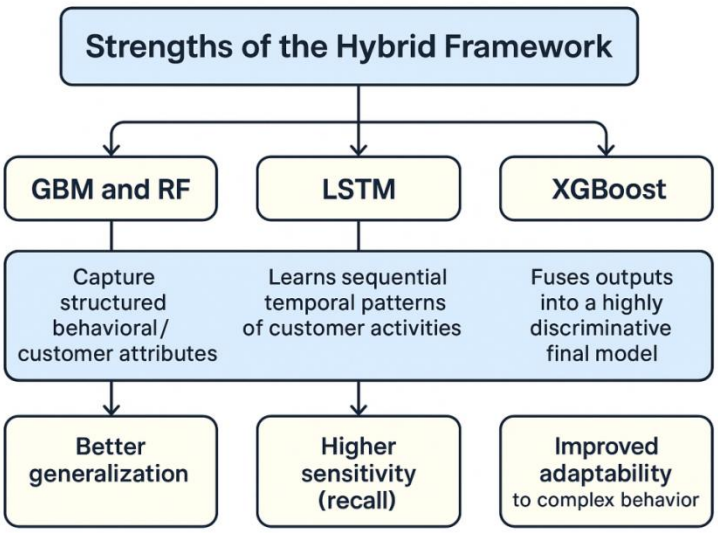

Figure 6. Strengths of the Hybrid Framework

Figure 6 is hybrid model of churn prediction uses three complementary predictive models: GBM/RF, LSTM, and XGBoost, to achieve high predictive accuracy. Random Forest and GBM are effective in capturing structured behavior and customer attributes hence good generalization. LSTM also leads to learning consecutive and time-based patterns of customer activity history, leading to greater recall and sensitivity. XGBoost is the meta-learner, which combines the performance of all base models into a high-level of discrimination final prediction. A combination of these methods leads to better generalization, sensitivity to churn signals and flexibility to advanced ways of customer behavior.

5.3 Comparative Analysis

- **LSTM**

The LSTM model achieved the greatest individual performance as it was a good model for capturing sequential dependencies in the data of customer activity.

- **GBM**

GBM helped to provide a good baseline by capturing non-linear interaction between structured customer features. This led to extraordinary performance with regard to tabular behavioral and demographic attributes.

- Random Forest (RF)

RF made stable and robust results by averaging the predictions over multiple decision trees.

While not so much sensitive to noise and overfitting, it lacked the fine-grained modelling ability of GBM and LSTM.Its performance was reliable and somewhat inferior to the other two individual models.

- **Stacking Ensemble**

The stacking ensemble exploited the power of GBM, RF, and LSTM using an XGBoost meta-learner. By using different patterns of learning from each base model, it was able to show superior generalization. This led to the highest AUC, which is better than all the standalone algorithms [20].

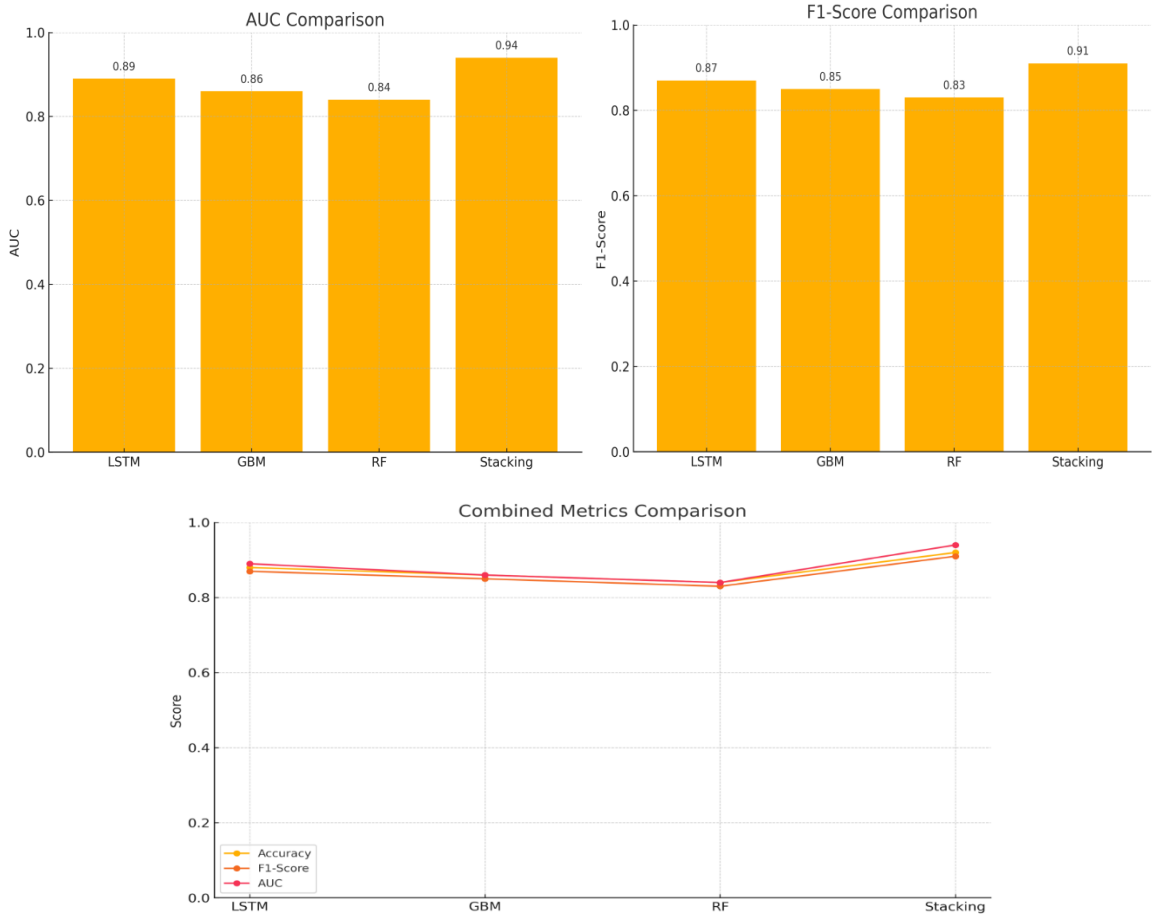

Figure 7(a), (b), (c) Combine AUC, F1

5. Conclusion

The Hybrid Ensemble Churn Prediction Framework is an efficacious integration of Gradient Boosting Machine (GBM), Random Forest (RF), and Long Short-Term Memory (LSTM) approaches to stacking methodological implementation reinforced by XGBoost. The simultaneous use of both systematic customer variables and sequence-based variables helps to significantly increase the overall precision, recall, and discriminatory value of the stacking methodological implementation to a considerable extent compared to other base approaches. The overall strong prediction power of this model helps online shopping platforms automatically identify customers and take corresponding actions to reduce business churn. The hybrid ensemble methodological implementation, although efficacious, emerges with multiple opportunities to enhance its capabilities. The model can be implemented in real-time by using micro services and/or streams. The use of Bayesian and/or evolutionary methods to optimize hyper-parameters helps to enhance precision while minimizing human involvement. The SHAP model helps to enhance understanding by comprehensively integrating business perspectives on systematic variables contributing to business churn to design more efficacious business interventions. The developed hybrid ensemble methodological implementation is effortlessly an efficacious, versatile technical tool to overcome business complexities to leverage its exemplary promise to more specifically design and implement more advanced personalization and automation processes.

6. Future Scope

The Hybrid Ensemble Churn Prediction Framework is a very strong base, but there are many ways to improve upon its capabilities. First, one can work on implementing it in real-time using stream processing technologies like Apache Kafka or cloud micro services to allow immediate notification for churn probability. Second, one can work on implementing Bayesian Optimization, Hyperband, and genetic algorithms to optimize hyper-parameters more effectively. Third, incorporating Explainable AI approaches like SHAP can increase and allow organizations to identify what behavioral factors significantly contribute to customers churning. Fourth, extending the dataset to incorporate clickstream activity, customer service engagement, and sentiment features could increase comprehensiveness. Fifth, one could work on more contemporary ideas like employing Deep Learn architectures like the Transformer, attention mechanisms, or TCN-Nets to improve long-term customer behavior understanding. Sixth, comparing implementation on multiple domains like finch, telecom, and retail can increase the applicability to more domains across which this tool can generally work well. Seventh, incorporating AutoML can automatically work on preprocessing; FS, and MS. Eighth, incorporating CSL and/or CLV can make predictions more aligned to business profitability. Ninth, one can work on implementing Federated Learning to train these ML approaches on customer data while being more secure. Finally, one can work on building interactive visualizing interfaces to increase organizational usability.

References

[1] X. Xiahou and Y. Harada, "B2C e-commerce customer churn prediction based on K-Means and SVM," *J. Theor. Appl. Electron. Commer. Res.*, vol. 17, no. 2, pp. 458–475, 2022.**DOI:** 10.3390/jtaer17020024

[2] M. I. et al., "Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning," *Mach. Learn. Knowl. Extr.*, vol. 7, no. 3, 2025. **DOI:** 10.3390/make7030105

[3] A. Bhatnagar and S. Srivastava, "Customer Churn Prediction: A Machine Learning Approach with Data Balancing for Telecom Industry," *Int. J. Computing*, vol. 24, no. 1, 2024.**DOI:** 10.47839/ijc.24.1.3873

[4] Md A. Al Rahib, N. Saha, R. Mia, and A. Sattar, "Customer data prediction and analysis in e-commerce using machine learning," *Bull. Electr. Eng. Inform.*, vol. 13, no. 4, 2024.**DOI:** 10.11591/eei.v13i4.6420

[5] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.**DOI:** 10.1214/aos/1013203451

[6] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001. **DOI:** 10.1023/A:1010933404324

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444,

2015.
DOI: 10.1038/nature14539

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *KDD*, pp. 785–794, 2016.
DOI: 10.1145/2939672.2939785

[9] W. Verbeke et al., "A comprehensive study on churn prediction in the telecommunications industry," *Expert Syst. Appl.*, 2011.**DOI:** 10.1016/j.eswa.2011.01.027

[10] M. Gupta and S. Tripathy, "Customer churn prediction in e-commerce: A machine learning approach," *Int. J. Data Sci. Anal.*, 2020.**DOI:** 10.1007/s41060-018-0158-0

[11] A. Amin et al., "Churn analysis using behavioral logs and usage patterns," *J. Ambient Intell. Humanized Comput.*, 2021.**DOI:** 10.1007/s12652-020-01855-4

[12] Z. Huang, H. Chen, and C. Zeng, "Applying data mining to telecom churn prediction," *IEEE ICDM*, 2018.
DOI: 10.1109/ICDM.2018.00015

[13] S. Zhang, A. Chen, and P. Zhou, "A hybrid ensemble model for customer churn prediction using stacking," *Procedia Comput. Sci.*, vol. 199, pp. 450–457, 2022.**DOI:** 10.1016/j.procs.2022.01.057

[14] A. Khattak et al., "A hybrid BiLSTM–CNN model for churn prediction," *Telecommun. Syst.*, 2023.
DOI: 10.1007/s11235-023-01060-w

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Oversampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.**DOI:** 10.1613/jair.953

[16] T. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible churn prediction models using rule induction," *Expert Syst. Appl.*, 2012.**DOI:** 10.1016/j.eswa.2011.10.027

[17] A. Idris, M. Rizwan, and A. Khan, "Churn prediction using boosting and bagging," *Eng. Appl. Artif. Intell.*, vol. 36, pp. 249–257, 2014.**DOI:** 10.1016/j.engappai.2014.07.001

[18] W. Wang et al., "Customer behavior prediction using deep learning," *IEEE Access*, 2018.
DOI: 10.1109/ACCESS.2018.2832981

[19] T. Fader and B. Hardie, "Customer-base analysis using Pareto/NBD models," *Marketing Science*, 2010.
DOI: 10.1287/mksc.1090.0505

[20] A. Graves, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
DOI: 10.1162/neco.1997.9.8.1735

[21] V. Kumar and S. Gupta, "Customer churn prediction using recurrent networks," *Information Systems Frontiers*, 2019.**DOI:** 10.1007/s10796-018-9865-3

[22] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, 2009.
DOI: 10.1109/TKDE.2008.239