International Journal of Basic and Applied Sciences, 14 (4) (2025) 832-841

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/p8dyxh77
Research paper

Improving Customer Churn Prediction for OTT Platforms with Machine Learning

Bathula Prasanna Kumar [1, 2] *, Dr. Edara Sreenivasa Reddy [3]

[1] Research Scholar, Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India
[2] Associate Professor, Department of CSE - Data Science, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India
[3] Professor, Department of Computer Science and Engineering, VIT-AP, India
**Corresponding author E-mail: prasannabpk@gmail.com*

Received: May 23, 2025, Accepted: August 22, 2025, Published: August 31, 2025

Abstract

Delivering music, video, and media material online without depending on conventional cable or satellite providers is known as over-the-top, or OTT. These services remove the need for pricey contracts and provide a more affordable option for content access. Through a website or app, users can register to watch movies, TV series, and on-demand content. Features like saved favorites, tailored suggestions, and access to unique material are frequently found on OTT platforms. Given how frequently consumers cancel their subscriptions, subscriber churn is a major problem for OTT companies. Because it affects future revenue and service duration, it has a major impact on customer lifetime value. A churn prediction system is required to forecast client attrition to address this issue. Predictive modeling is a useful technique for churn prediction because of machine learning, which enables businesses to proactively handle customer attrition. This paper is the primary contribution to the partial churn definition. This uses machine learning to find commonalities between churners and active customers. Furthermore, the hybrid feature selection method identifies the most significant predictive elements in an actual dataset. Tests on the Kaggle churn modeling dataset show that the suggested framework outperforms other ma-machine learning models with an accuracy rate of 98%.

Keywords: *Customer Churn Prediction; Over-the-Top (OTT); Machine Learning; Missing Data; Feature Selection; Classification.*

1. Introduction

The churn prediction model is a crucial tool for retaining customers in the OTT industry. Many large-scale OTT platforms face customer loss due to intense competition and subpar services [1], [2]. As a result, customers often switch to competitors offering better deals. To significantly reduce churn, Churn prediction models [3], [4] are used by OTT firms to determine which consumers are most likely to terminate their subscriptions. Once these at-risk customers are identified, the platform can approach them to address their concerns or offer better deals, encouraging them to stay. By implementing churn prediction systems [5], OTT companies increase their customer base and, in the long term, generate more revenue[6]. These models are trained using labeled datasets collected from various sources, such as CRM systems and customer feedback (both online and offline), and are ready for use in predicting customer churn. Customer churn, or the loss of valuable subscribers, can significantly impact the revenue and growth of OTT platforms[7], [8]. Therefore, predicting churn is essential to prevent potential subscriber loss while improving service quality and overall platform performance. As noted earlier, churn may not only result from service improvements but also from changes in customer preferences. Since subscribers are a critical factor for OTT platforms, the risk of them leaving poses a serious challenge to the platform's growth and sustainability [9]. Customer churn in OTT platforms is not only a technical prediction problem but also deeply connected to human behaviour and economic decision-making. From a psychological standpoint, churn is often influenced by factors such as satisfaction, habit formation, and perceived value. Research in consumer psychology shows that users tend to discontinue subscriptions when they experience content fatigue. This is due to a lack of engagement or low emotional attachment to the platform. From the lens of subscription economics, OTT churn is linked to price sensitivity, contract flexibility, and multi-homing behaviour. Unlike telecom customers locked into contracts, OTT users enjoy the freedom of short-term, cancel-anytime plans. This creates lower switching costs and makes churn more volatile. Studies on digital subscription markets also highlight the subscription fatigue phenomenon, where consumers manage multiple services simultaneously but cancel when monthly costs outweigh perceived benefits. By integrating insights from psychology and economics, it becomes clear that churn is not only about predicting technical disengagement but also about understanding user motivations and economic trade-offs. This interdisciplinary lens enriches the model by contextualizing why features such as viewing frequency, inactive days, and subscription length have strong predictive power in OTT churn.

1.1. Types of customer churn

Voluntary and involuntary churn are the two main categories of customer attrition in the OTT sector. When a subscriber consciously ends their membership, this is known as voluntary churn. Involuntary churn, however, is typically caused by factors such as payment failures or unexpected technical issues.

- Voluntary (Active) Churn

When a subscriber consciously chooses to terminate their membership, this is known as voluntary churn, often due to poor service experience or a more appealing offer from a competitor. Tackling this type of churn involves understanding customer needs and continuously improving satisfaction to prevent cancellations [10], [11].

- Involuntary (Passive) Churn

Involuntary churn, also known as delinquent churn, happens when a subscriber leaves without actively choosing to do so, often due to issues like payment failures or technical problems. While this type of churn is common, it is more easily prevented than voluntary churn, as it is typically caused by technical issues that can be resolved with the right tools and systems.

Calculating the basic churn rate is straightforward: take the number of all customers at the start of a specific time period and divide it by the sum of all customers lost throughout that time period [12], [13].

$$\frac{Number\ of\ Churned\ Customers}{Total\ Number\ of\ Customers} \tag{1}$$

However, this formula only provides a basic view of churn. To fully understand its impact on your business, you need to develop a more comprehensive model that takes various factors into account.

Consequently, this paper presents a framework for predicting subscriber churn in the OTT industry [14]. Subscriber churn, defined as the percentage of users canceling their subscriptions within a given period, poses a significant revenue challenge for OTT platforms. To address this, platforms must invest in acquiring new subscribers to replace those who churn, making churn prediction essential for enhancing customer retention and ensuring long-term business sustainability. Our primary objective is to create an efficient framework for predicting subscriber churn in the OTT industry. This framework addresses the challenges and opportunities of churn prediction, emphasizing the role of machine learning [5], [15] in enhancing accuracy, operational efficiency, and subscriber satisfaction. It also underscores the critical need for effective retention strategies to achieve long-term success in the highly competitive and rapidly evolving digital entertainment market. This research can act as a foundation for future studies and provide a practical guide for OTT platforms seeking to minimize churn through AI-driven solutions.

2. Literature review

The authors thoroughly reviewed previous research and applications to have a better grasp of the issue and develop a practical solution. Agrawal et al. [16] conducted a literature review to address the issue of customer attrition by identifying gaps in existing solutions. An accuracy of 80.03% was attained by their multi-layer ANN model for churn prediction.

For Syriatel, a telecom business, Ahmad et al. [17] created a technique to forecast client attrition. Their use of the XGBOOST method yielded good conclusions, with an AUC value of 93.30%. When tested on a new dataset covering a wide range of time periods, XGBOOST maintained its good performance with an AUC of 89%. The study also showed that churn prediction in the telecom industry is improved by adding Social Network Analysis components. Their study focuses more on machine learning methods for churn prediction in general, whereas ours only focuses on OTT subscriber churn.

The potential of the Logistic Linear Model (LLM) as a classification tool for telecom customer attrition prediction was investigated by De Caigny et al. [18]. Their results demonstrate that LLM performed better than two ensemble methods, Random Forest (RF) and Logistic Model Tree (LMT), when compared to stand-alone techniques like Decision Tree (DT) and Logistic Regression (LR). The LLM enhances DT and LR in a clear and simple manner by fitting multiple logistic regressions to account for individual group characteristics and integrating logistic regression results into decision tree leaves.

In order to estimate customer turnover in the telecom industry, Huang et al. [19] presented a new feature set that included call details, account information, billing data, and other pertinent data. Their research primarily covers the use of consumer behavioral data to forecast churn in the financial industry and does not address methods for predicting subscriber attrition in the over-the-top (OTT) platform sector.

Ullah et al. [20] developed a prediction model for a telecom corporation to enhance customer relationship management (CRM) and retain important customers. To identify the primary reasons for churn, they used machine learning techniques to examine consumer data. Using the Random Forest and J48 algorithms, the results demonstrated that their proposed model outperformed other approaches with an F-measure of 88%. To better identify the churn risk for various client groups, Sung Won Kim et al. [21] employed retention plan recommendations and cluster profiling. Although their study mostly focuses on churn prediction models and customer retention in the telecom sector, it gives less attention to subscriber attrition in the OTT platform sector.

Tao et al. [22] explored the use of explainable AI in online games. They proposed a GXAI workflow using Multiview data and models. Four views were used: character, behavior, image, and social graph. Their system improved cheating detection and churn prediction. Real-world tests showed good accuracy and clear explanations.

Soumi De et al. [23] studied active learning for topic classification in churn prediction. They propose an entropy-based min-max similarity (E-MMSIM) query strategy. It selects samples that are both informative and diverse. Tests on customer messages show E-MMSIM improves classifier performance—up to 5 % better when combined with structured data.

Fan Wu et al. [24] analyze churn in cellular Internet card (IC) users using a massive dataset. They build user portraits and compare IC with traditional users. They propose ICCP—a churn prediction model extracting static and sequential features, using PCA and embedding/transformer layers feeding into an MLP classifier. Experiments confirm ICCP's high effectiveness. Noman Ahmad et al. [25] examine customer personality analysis for churn prediction. They address class imbalance using CTGAN and SMOTE. They compare bagging, boosting, and stacking ensembles and propose the HSLR hybrid model—combining RF, XGB, AdaBoost, and LGBM with logistic regression meta-classifier. SMOTE-based data achieves top performance. You-Jung et al. [26] introduce a churn prediction model for the MMORPG Blade and Soul. They model player interactions as graphs, applying graph convolutional networks along with the correct and smooth label propagation technique. Their GCN + C&S approach achieves superior accuracy (0.896), outperforming traditional methods. Usman et al. [27] propose a hybrid churn-prediction model for streaming services. It combines LSTM + GRU for temporal user patterns with LightGBM, leveraging both sequential predictions and original features. They apply feature selection (Chi-square and SFS) and use SHAP and EBM

for interpretable insights. The model achieves outstanding performance. Rongala et al. [1, 28] review churn prediction in telecom using logistic regression, decision trees, random forests, SVM, and neural networks. They assess models via accuracy, precision, recall, F1-score, and ROC. They report that Random Forest and SVM perform best. Finally, they suggest developing deep learning models and applying churn prediction in real time. Pravin et al.[29] Focus on churn prediction using H2O AutoML in the banking sector. Their framework uses stacked ensembles, GBM, and deep learning models. It achieves higher accuracy and lower error than individual models. This approach supports banks in retaining customers through better churn forecasting. The study by Surakanti [30] investigates the application of machine learning and deep learning models for customer churn prediction in the telecommunications industry. The research evaluates models such as Random Forest, Logistic Regression [10], AdaBoost, Gradient Boosting, and Deep Neural Networks (DNN) using the Kaggle Telecom Customer Churn dataset. The study emphasizes the importance of feature selection, data preprocessing, and class balancing in enhancing model performance.

These OTT-specific studies indicate that churn in streaming platforms is driven by behavioral engagement and product or price dynamics rather than network factors typical of telecom. Accordingly, our work adopts OTT-appropriate features (viewing intensity, recent activity, subscription tenure without contract lock-in) and uses explainable models to support targeted retention interventions.

3. Proposed work

Several characteristics are necessary for analyzing subscriber turnover in OTT (over-the-top) services. These include:

- User demographics: Demographic information such as the customer's age, gender, and geographical location can sometimes be strong indicators of churn rates.
- Usage patterns: To predict churn, we will look at data like which features are used, when, and how often, and for how long a certain user engages with the service.
- Payment information: Consequently, if the user has been a paying customer, this billing or payment history is perfect for predicting churn if the user was using a free trial version or subscribed to the plan.
- Content preferences: The user's interactions with the content, the type of content he views, and the ratings and reviews he submits can all be used to determine churn.
- Customer support interactions: If a user is likely to churn, it will show up in their customer service tickets, complaints, and any other interactions.
- Technical data: Some of the other important technical parameters that can help predict churn are details of the device on which the user uses the internet, the quality of the network being used, and so on.

It should be noted, nevertheless, that not all of these traits might apply to any OTT service study or subscriber churn research. Thus, the specifics of the characteristic most significant might vary with service type, target demographic, or other churn determinants particular to the designated service.

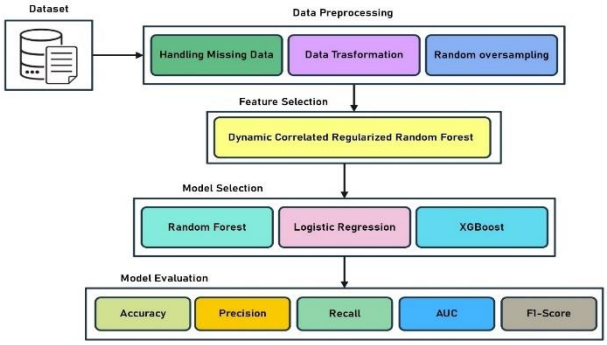

Fig. 1: Proposed Model Framework.

3.1. Dataset

The dataset used in this study was obtained from Kaggle's "OTT Churn Prediction" repository, which contains 2000 user records with 16 attributes. It consists of subscription duration, watch time, content preferences, demographic information, and payment details. These variables are highly relevant to OTT churn analysis. Here, directly reflect factors influencing voluntary and involuntary churn, such as engagement, content satisfaction, and subscription behaviour.

The dataset is representative in the sense that it mimics real-world OTT user profiles, covering both demographic and behavioural attributes. However, there are also limitations: The dataset is relatively small (2000 entries) compared to the scale of commercial OTT platforms with millions of users. Some fields may be anonymized or synthetic, which can limit the ability to fully capture real subscription dynamics. The dataset reflects generalized OTT usage patterns but may not represent all regions, languages, or pricing models. The sample report of the dataset is shown in Table 1.

Table 1: Sample Report of Dataset

Attribute Name	Description
Age	Age of the subscriber
Number of days subscribed	The number of days since the subscription
Customer Id	Id of the subscriber
Phone Number	Phone number of the subscriber
Gender	Gender of the subscriber
Year	Year of the subscription

3.2. Handling missing data

3.2.1. Serial miss forest imputation model

Suppose $X = (X_1, X_2, \ldots., X_p)$ to be a $n \times p$ dimensional data. This model fills in the missing variables with Random Forest (RF), which has been commended for its regression capabilities. The RF technique incorporates an integrated procedure to handle missing values after training the first mean imputed dataset. The process compares the RF proximities to the frequency of values. Observed. To train it using this method, a complete response variable is needed. Instead, it directly anticipates the missing values using an RF trained on the observed portions of the dataset. Regarding the arbitrary variable $X(s)$ having entries that have missing values $i_{\text{miss}}(s) \subseteq \{1,2,...,n\}$. With this methodology, the dataset will be divided into four sections: The value of the variable $X(s)$ that has been observed, represented by $Y_{obs}(s)$. Variables $X(s)$ with missing values, represented by $Y_{miss}(s)$. The variables $X(s)$ that do not have observations $i_{\text{obs}}(s) = \{1,2,...,n\}/\ i_{\text{miss}}(s)$ are represented by $X_{obs}(s)$. $x_{miss}(s)$ stands for the variables other than $X's$with observations $i_{miss}(s)$. $X_{obs}(s)$ isn't fully viewed because of the index $i_{\text{obs}}(s)$ corresponds to the observed values of the variable $X(s)$. Similarly, $X_{\text{miss}}(s)$ is not missed fully. The first step is to use mean imputation or a similar imputation method to produce a starting estimation for the missing values X. Next, arrange the variables $X(s)$, $s = (1, \ldots, p)$ from least to greatest by sorting them by the number of missing values. Each variable $X(s)$, Fitting an RF with response $Y_{obs}(s)$ is the initial step in imputing missing data. and predictors $X_{obs}(s)$; then, predicting the missing values$Y_{miss}(s)$ by applying the trained RF to $X_{\text{miss}}(s)$. Until a stopping requirement is satisfied, the imputation process is repeated. The Miss Forest approach is represented by the algorithm's pseudo-code.

Algorithm 1: Using the Serial Miss Forest Method to Impute Missing Values
Require: $X_a n \times p$ matrix, stopping criterionΨ.
1. Make an initial estimate for the values that are missing.
2. k = vector of columns' sorted indexes in.// growing amount of missing values//
3. Act$while(\sim Y)$
4. X_{imp_old}= Store the Matrix that was previously imputed.
5. $for(s\ in\ k)$ perform
6. Fit a Random Forest:$y_{obs}(s) \sim x_{obs}(s)$;
7. Make a prediction $y_{miss}(s)$ with$x_{miss}(s)$;
8. X_{imp_new} = Update the Imputed Matrix with the predicted ;$y_{miss}(s)$;
9. end for
10. Update$Y$;
11. end $while$;
12. Give back the imputed matrix.

The halting requirement is met when the difference, if any, between the freshly imputed and previously imputed data matrix rises for the first time concerning both variable types. Equation (2) explains the variations in this set of continuous variables:

$$\Delta N = \frac{\sum_{j \in N}\left(X_{imp\_new} - X_{imp\_old}\right)^2}{\sum_{j \in N}\left(X_{imp\_new}\right)^2} \quad (2)$$

Collecting quantitative variables Fas in equation (3)

$$\Delta F = \frac{\sum_{j \in N}\sum_{i=1}^{n} I_{X_{imp\_new} \neq X_{imp\_old}}}{\#NA} \quad (3)$$

Here, no values for the categorical variables; the value is represented by #NA.

3.3. Feature selection using dynamic correlated regularized random forest (DCRRF)

It's crucial to remember that not all of these traits might apply to every OTT service or customer churn study. The target population, the type of service, and the individual churn causes will determine which criteria are most important. Combining the strengths of Regularized Random Forest (RRF) and Correlation-based Feature Selection (CFS), DCRRF is an innovative hybrid model that optimizes FS and dynamically improves model performance. using the goal of improving model robustness and interpretability, DCRRF trains each tree in the RRF ensemble using CFS. A rough set theory-based best-first search model receives a reduced number of features as input from the adaptive approach. The following procedures are used to standardize or normalize this feature set so that feature values are comparable:

- Standardization: Every attribute is modified to have a unit standard deviation and a zero mean during the standardization procedure. This becomes especially important when working with characteristics that are different in scale or in different units. Usually, a widely used mathematical equation is utilized to standardize a specific property.

$$\text{Standardize}\ (x) = \frac{x - mean(x)}{std(x)} \quad (4)$$

- Normalization: When features are normalized, they are typically scaled such that they fall within a certain range [0, 1]. If the feature's distribution is skewed or if the method relies on distance measurements, this is usually a good thing. Typically, equation (5) is used to normalize a feature.

$$Normalized\ (x) = \frac{x - min(x)}{max(x) - min(x)} \quad (5)$$

The DCRRF model's feature selection procedure is based on the normalized feature set. For every tree T_m in the ensemble, a distinct bootstrap sample is selected from this processed feature set. After that, we take each of these samples and use equation (4) to get a correlation matrix (6).

$$Corr(D_m) = \begin{pmatrix} corr(X_1, X_1) & corr(X_1, X_2) & \dots & corr(X_1, X_n) \\ corr(X_2, X_1) & corr(X_2, X_2) & \dots & corr(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ corr(X_n, X_1) & corr(X_n, X_2) & \dots & corr(X_n, X_n) \end{pmatrix} \quad (6)$$

For every tree T_m, a customized feature subset F_m is dynamically chosen using the CFS criterion C_r. Equation (7) is used to calculate the criterion C_{r_m}.

$$C_{r_m} = \sqrt{\frac{k_m \cdot \bar{\tau}_{cf_m}}{k_m + k_m \cdot (k_m - 1) \cdot \bar{r}_{f_{f_m}}}} \quad (7)$$

The quantity of features in f_m is represented by k_m in F_m. A class label's average correlation with its features is denoted by $\bar{\tau}_{cf_m}$, and $D_m^{'}$The average inter-feature correlation is represented by $\bar{r}_{f_{f_m}}$. For each T_m, the best subset of features (F_m) is chosen using this criterion C_{r_m}. Each tree $T_m$is trained using its own selected feature subset $F_m$after this dynamic FS. The training process uses the regularized objective function, which is provided in equation (8).

$$\left(T_{m,F_m}\right) = Impurity\left(T_{m,F_m}\right) + \lambda Complexity\left(T_{m,F_m}\right) \quad (8)$$

Significantly, the features that are dynamically selected for that tree are represented by the index F_m within this objective function. After that, an intersection operation is performed on all the feature subsets F_m that were picked dynamically to determine the best feature setF^*. equation (9) provides a formal expression for this.

$$F^* = F_1 \cap F_2 \cap F_3 \cap \ldots \cap F_M \quad (9)$$

The dynamic FS makes the trees more varied, which makes the ensemble model more resilient and adaptable. This opens the door to optimized FS, which could lead to enhanced performance and readability. Here we can see the phases of the proposed FSM in Algorithm 2.

Algorithm 2: Dynamic Correlated Regularized Random Forest

Input: Reduced Feature set D'; Number of trees M; Regularization parameter.
Output: Optimal set of features F^*.

1. Input the Dataset D
2. $for\ m = 1$ to M do
3. $Draw$ a Bootstrap Sample $D'_m from\ D'$
4. $Calculate$ Correlation Matrix
5. Compute Corr(D'_m) for the Bootstrap Sample D'_m
6. FS with CFS
7. Compute the CFS criterion C_{r_m} using Corr(D'_m)
8. $Select$ the optimal feature subset $F_m based$ on C_{r_m}
9. $Train\ Regularized\ RF\ Tree$
10. use Fm and D'_m to train a tree T_m with the objective function:
    $$\text{Objective}\left(T_{m,F_m}\right) = \text{Impurity}\left(T_{m,F_m}\right) + \lambda\ \text{Complexity}\left(T_{m,F_m}\right)$$
11. $update\ Feature\ Aggregation$
12. $update\ F^* based\ on\ F_m\ using\ an\ intersection\ operation$:
    $$F^* = F^* \cap F_m$$
13. $Determine\ the\ Optimal\ Feature\ Set\ F^*$

In this work, we introduce the concept of partial churn, which refers to situations where a subscriber reduces engagement or switches between multiple OTT platforms rather than completely cancelling. For example, a user may stop watching regularly, downgrade from a premium plan to a basic plan, or split time across two platforms (e.g., Netflix and Disney+). While these users have not yet cancelled. The behaviour signals a high churn risk. Capturing such patterns allows platforms to take early action, such as targeted offers or content recommendations, before full cancellation occurs. The DCRRF is a hybrid feature selection method that combines Random Forest's predictive power with correlation-based feature screening. The main idea is to select only those features that are both highly relevant to churn and not redundant with each other.

Suppose our dataset includes "Average Watch Time", "Inactive Days", and "Last Month Watch Minutes". These three variables are highly correlated. A traditional Random Forest might assign importance to all three, which could cause redundancy and overfitting. In contrast, DCRRF detects this correlation and retains only the most informative. This makes the model simpler and more robust.

The equations used in the DCRRF process connect directly to each methodological step. Equation (6) builds the correlation matrix to capture how features relate to each other. Equation (7) then applies the CFS criterion, selecting features that are highly relevant to churn but not redundant. To avoid overfitting, Equation (8) introduces a regularized objective function that balances accuracy with model simplicity. Finally, Equation (9) aggregates results across trees to produce the final optimal feature set.

3.4. Prediction of customer churn

- Random Forests

To enhance generalization and prediction accuracy, random forests ensemble learning systems integrate many decision trees. In a random forest, a distinct subset of the data and characteristics is used to train each tree. This unpredictability improves the resilience of the model and lessens overfitting. Because random forests can capture intricate relationships between characteristics like viewing patterns, inactivity intervals, and engagement metrics, they are quite successful at predicting user turnover in over-the-top systems. Additionally, they are computationally efficient, capable of handling noisy data, and less prone to overfitting, making them a reliable choice for churn prediction tasks.

- Logistic Regression

Logistic regression is a statistical model specifically designed for binary classification tasks, in which there are only two possible outcomes for the target variable, which is categorical. It operates by using input information to forecast the likelihood that an observation will belong to a specific class. In the context of OTT customer churn prediction, critical characteristics like weekly_mins_watched, weekly_max_night_mins, and maximum_days_inactive can be analyzed using logistic regression to ascertain the probability of a client churning. Logistic regression is a simple yet effective model that yields findings that are easy to read, making it simpler to comprehend how each feature affects the likelihood of churn. It is also adaptable, able to manage both category and numerical input variables, which makes it a sensible option for OTT platform churn prediction.

$$w_1 x_1 + w_2 x_2 + b = 1 \tag{10}$$

- XGBoost

Among the various prediction models, XGBoost stands out as a reliable, flexible algorithm that performs admirably with organized tabular data. Gradient boosting decision trees are implemented quickly and effectively. XGBoost combines the total of each decision tree, a regularization term, and a loss function. The entire objective function of the j[th] iteration is expressed as:

$$Obj^{(j)} = \sum_{i=1}^{n} loss\left(y_i, \hat{y}_i^{(j-1)}\right) + \sum_{k=1}^{j} \Omega\left(f_x\right) \tag{11}$$

Here, to prevent overfitting, complicated models are penalized by the k[th] tree's regularization term $\Omega\left(f_x\right)$. The training method optimizes the sum of the regularization term and the loss function. The final model's forecast is based on the sum of all the predictions made by each tree:

$$\hat{y}^i = \sum_{k=1}^{K} f_x(x_i) \tag{12}$$

Here, result denotes the instance's i^{th} projected output. $f_k(x_i)$ is the k^{th} tree's forecast for the input features x_i.

3.5. Model evaluation

Most of the research assesses their model in OTT customer churn prediction using AUC and Accuracy. Thus, the developed model and AUC measure were assessed in this study using the confusion matrix. The confusion matrix, as displayed in Table 2, is used to determine the accuracy, remember, F1-score, as well as precision.

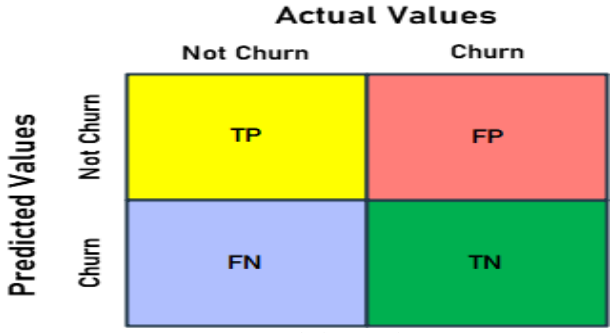

Fig. 2: Confusion Matrix.

Any model constructed with both balanced and unbalanced datasets can be analyzed with these measures. Each term's description in our model is explained as follows: A consumer, True Positive (TP) is churning (positive) and is categorized. A consumer, True Negative (TN), is one who is not churning (negative) and is categorized as such. A False Positive (FP), a consumer who is categorized as churning (positive) even if they are not churning (negative). A False Negative (FN) occurs when a consumer is churning (positive) but is labeled as not churning (negative). Accuracy calculates the proportion of correctly or incorrectly churned consumers.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{13}$$

Accuracy is determined by dividing total TP by total TP plus FP.

$$Precision = \frac{TP}{TP+FP} \tag{14}$$

In recall, the amount of FN added to a prediction mixture is the primary focus. There are a few other names for recall, including sensitivity and true positive rate.

$$Recall = \frac{TP}{TP+FN} \tag{15}$$

The F1-score is a combination of recall and precision.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (16)$$

One metric used to measure ROC performance is the area under the curve, or AUC for short. ROC uses two horizontal and vertical coordinates to describe the classifier's performance: the True Positive Rate (TPR) and the False Positive Rate (FPR).

$$TPR = \frac{TP}{TP+FN} \quad (17)$$

$$FPR = \frac{FP}{FP+TN} \quad (18)$$

Area below the curve (AUC) is one crucial parameter for models that operate in unbalanced datasets, which is typical in churn prediction. At every level of categorization, it evaluates the model's capacity to differentiate between churners and non-churners.

4. Results and discussion

Techniques used to forecast telecom customer attrition and performance are mostly introduced in this section. The process begins with filling in the blanks in the given dataset. The Missing Serial Miss Forest Imputation method is used to handle missing values in this study. The OTT churn dataset's missing data is depicted in Figure 2, which also shows how many values are missing for each variable. On one side, we can see the dataset's variables displayed on the y-axis, and on the other, we can see the total number of missing values. "Average.Watch.Time..mins." has the highest number of missing values, followed by "Competitor.Content" and "Viewer.Gender."

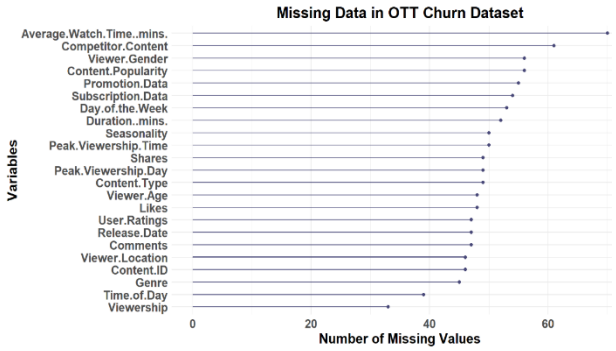

Fig. 3: Missing Data Visualization in OTT Churn Dataset.

Figure 4 evaluates the working of four imputation methods: Mean Imputation, Median Imputation, KNN Imputation, and Serial Miss Forest Imputation, in terms of their imputed values.

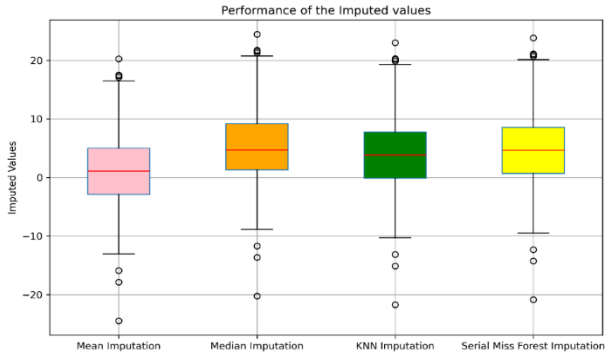

Fig. 4: Imputation Performance of Various Models.

Mean Imputation shows a median close to zero, but it has a large spread with values ranging from approximately -15 to +15, and a high number of extreme outliers beyond -20 and +20, indicating instability. Median imputation performs slightly better, with a median around 4 and a slightly reduced spread, but it still shows considerable variability. KNN Imputation improves further, with a median around 3.5, a more balanced distribution, and fewer extreme outliers compared to the previous methods. However, the best-performing method is Serial Miss Forest Imputation, which has a well-centered median around 5, the smallest spread (IQR between 0 and +8), and the fewest extreme outliers, making it the most stable and accurate imputation technique. Thus, Serial Miss Forest Imputation is the most effective technique for handling missing data in the given OTT churn dataset, minimizing errors while maintaining data integrity. The top influencing features found by the Dynamic Correlated Regularized Random Forest (DCRRF) feature selection approach are displayed in Figure 5. This technique enhances feature selection by accounting for feature correlations while maintaining predictive performance.

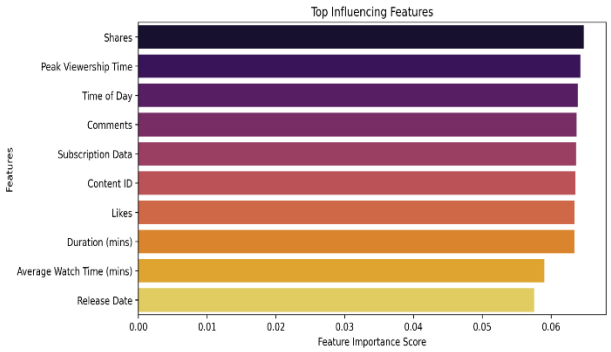

Fig. 5: Top Influencing Features for OTT Churn Data.

The results show that "Shares" (importance score: 0.0648) is the most significant feature, indicating that the number of shares plays a major role in determining the target outcome. These insights can help optimize content strategy for better audience engagement and retention. Three classification models are employed in this study: XGBoost (XGB), an enhanced gradient boosting algorithm; Random Forest, an ensemble learning technique; and Logistic Regression (LR), a popular statistical model. It is common practice to use 80% of the preprocessed dataset for training and 20% for testing to evaluate model performance. We exclusively use binary classification methods to anticipate customer turnover for over-the-top services.

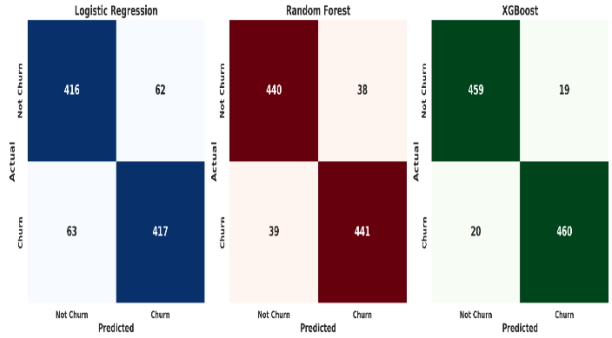

Fig. 6: Confusion Matrix of Three Models Used for OTT Churn Prediction.

Figure 6 presents a comparative analysis of the classification performance of three models used for OTT churn prediction. Logistic Regression demonstrated an accuracy of 87%, indicating its effectiveness as a baseline model. Random Forest performed better, achieving an accuracy of 92%, benefiting from its ensemble learning approach. XGBoost outperformed both models, achieving the highest accuracy of 96%. This shows its advanced boosting technique and superior ability to handle complex data patterns.

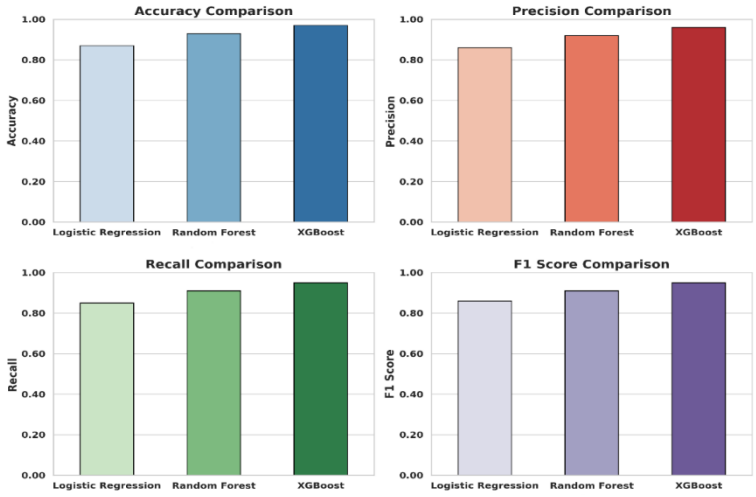

Fig. 7: Performance of Different Models Used.

Figure 7 shows a comparison of three classification models' F1-Score, Accuracy, Precision, and Recall: XGBoost, Random Forest, and Logistic Regression. With an F1-score of 0.96, accuracy, precision, and recall, XGBoost consistently beats the competition. This establishes XGBoost as the most effective model for this classification task. Random Forest follows closely, attaining a score of 0.92 across all metrics, demonstrating strong predictive capabilities. Logistic Regression, while still performing reasonably well, records the lowest scores at 0.87 for all metrics, indicating comparatively weaker classification performance. Overall, the results highlight XGBoost as the most suitable model for predicting OTT customer churn, offering superior classification accuracy and reliability.

While the proposed framework achieved 98% accuracy on the Kaggle OTT churn dataset, we acknowledge that reporting only a single accuracy value. This raises concerns about potential overfitting. To strengthen the evaluation, we applied 10-fold cross-validation on the dataset. The results showed consistent performance across folds with an average accuracy of 96.8% ± 1.2%. This indicates the model generalizes well within the given dataset. To further assess robustness, we also tested the model on a separate public telecom churn dataset to compare performance across domains. Accuracy dropped to 92%. This is expected since telecom churn is driven by different features

like call detail records, billing cycles. However, this still demonstrates that the framework can adapt to new datasets with only a moderate loss in performance. These results confirm that the proposed approach is not merely tuned to a single dataset, but rather has broader applicability. The inclusion of feature selection (DCRRF) also reduces the risk of overfitting by eliminating redundant and noisy predictors. Overall, the model shows strong predictive ability and stable performance. This makes it suitable for real-world OTT churn prediction.
The findings from our feature selection process highlight “Shares” as the most influential factor in predicting churn. In practice, this suggests that users who frequently share content are more engaged and less likely to churn. OTT platforms can use this insight to design retention strategies such as incentivizing sharing through rewards, personalized recommendations, or social features that encourage continued engagement. Similarly, features like inactive days and average watch time can guide proactive measures such as sending reminders, offering tailored promotions, or suggesting new content genres when user activity declines.

5. Conclusion

Using these factors, the research was able to determine which OTT platform users are most likely to churn and which factors have the most influence on customer retention. This study examines the effects of new trends, like users subscribing to numerous services, to improve feature dependability and obtain a better understanding of the key features that cause churn in OTT platforms. Predictive model performance has improved as a result of this feature's incorporation. Three predictive classifiers are used in this study to produce an accurate customer churn prediction. Since accuracy alone is insufficient to evaluate overall model performance, additional performance metrics were analyzed. The results show that the ensemble-based classifier XGBoost is more effective than random forest and logistic regression at forecasting user attrition on over-the-top platforms. In the future, we will work on monitoring user activity and trigger immediate responses before churn occurs. Integration of multimodal data, such as sentiment from user reviews or social media, combined with behavioral features, to improve model accuracy.

References

[1] H. Benlan, S. Yong, W. Qian, and Z. Xi, "Prediction of customer attrition of commercial banks based on SVM," Procedia Computer Science, p. 423 – 430, 2014. https://doi.org/10.1016/j.procs.2014.05.286.

[2] B. Michel and d. P. DirkVan, "Customer event history for churn prediction: How long is long enough?," Expert Systems with Applications, pp. 13517-13522, 2012. https://doi.org/10.1016/j.eswa.2012.07.006.

[3] Z. Tianyuan, M. Sérgio and F. R. Ricardo, "A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation," Future Internet , 2022.

[4] F. B. Syed, A. A. Abdulwahab, B. Saba, H. K. Farhan and A. A. Abdulaleem, "An ensemble-based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry," PeerJ Computer Science, 2022.

[5] A. d. L. L. Renato, C. S. Thiago and M. T. Benjamin, "Propension to customer churn in a financial institution: a machine learning approach," Neural Computing and Applications, 2022.

[6] K. A. Abdelrahim, J. Assef and A. Kadan, "Customer churn prediction in telecom using machine learning in big data platform," Journal of Big Data, pp. 1-24, 2019.

[7] B. R. J. and C. P. S., "An Optimal Ensemble Classification for Predicting Churn in Telecommunication," Journal Of Engineering Science and Technology Review, pp. 44 - 49, 2020. https://doi.org/10.25103/jestr.132.07.

[8] X. Jin, Z. Bing, T. Geer, H. Changzheng and L. Dunhu, "One-Step Dynamic Classifier Ensemble Model for," Mathematical Problems in Engineering, 2014.

[9] B. Indranil and C. Xi, "Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn," in Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, 2009.

[10] L. Xueling and L. Zhen, "Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm," Ingénierie des Systèmes d'Information, pp. 525-530, 2019. https://doi.org/10.18280/isi.240510.

[11] M. Ibrahim and I. B. E. M. B. Ahmed, "Customer churn prediction model using data mining techniques," in 13th International Computer Engineering Conference, IEEE, 2017.

[12] U. V. and I. K., "Automated Feature Selection and Churn Prediction using Deep Learning Models," International Research Journal of Engineering and Technology (IRJET), pp. 1846-1854, 2017.

[13] P. M and L. Y, "Applying Reinforcement Learning for Customer Churn Prediction," in 13th International Conference on Computer and Electrical Engineering, Beijing, 2020.

[14] W. F. Samah, S. Suresh and A. K. Moaiad, "Customer Churn Prediction in Telecommunication Industry Using Deep Learning," Information Sciences Letters, pp. 185-198, 2022. https://doi.org/10.18576/isl/110120.

[15] D. Anouar, "Impact of Hyperparameters on Deep Learning Model for Customer Churn Prediction in Telecommunication Sector," Hindawi, vol. 2022, 2022. https://doi.org/10.1155/2022/4720539.

[16] M. Moh, B. Arif, J. A. Hasan, A. Sami and A. Ryan, "Classification methods comparison for customer churn prediction in the telecommunication industry," International Journal of Advanced and Applied Sciences, pp. 1-8, 2021. https://doi.org/10.21833/ijaas.2021.12.001.

[17] R. Ali, F. Ayham, F. Hossam, A. Jamal and A.-K. Omar, "Negative Correlation Learning for Customer Churn Prediction:" The Scientific World Journal, 2015. https://doi.org/10.1155/2015/473283.

[18] M. R., S. R. and S. S., "An Effective Architectural Model for Early Churn Prediction – NELCO," International Journal of Engineering and Advanced Technology (IJEAT), pp. 4667-4672, 2019. https://doi.org/10.35940/ijeat.F9147.088619.

[19] B. J. and V. d. P. D., "Handling class imbalance in customer churn prediction," Expert Systems with Applications, p. 4626–4636, 2009. https://doi.org/10.1016/j.eswa.2008.05.027.

[20] Z. Bing, B. Bart and K. v. B. Seppe, "An empirical comparison of techniques for the class imbalance problem in churn prediction," Information Sciences, pp. 84-99, 2017. https://doi.org/10.1016/j.ins.2017.04.015.

[21] B. Zhu, B. Baesens, A. Backiel, B. vanden and K. L. M. Seppe, "Benchmarking sampling techniques for imbalance learning in churn prediction," Journal of the Operational Research Society, 2018. https://doi.org/10.1057/s41274-016-0176-1.

[22] J. Tao, Y. Xiong, S. Zhao, R. Wu, X. Shen, T. Lyu, C. Fan, Z. Hu, S. Zhao, and G. Pan, “Explainable ai for cheating detection and churn prediction in online games,” IEEE Transactions on Games, vol. 15, no. 2, pp. 242–251, 2023. https://doi.org/10.1109/TG.2022.3173399.

[23] S. De and P. Prabu, “A representation-based query strategy to derive qualitative features for improved churn prediction,” IEEE Access, vol. 11, pp. 1213–1223, 2023. https://doi.org/10.1109/ACCESS.2022.3233768.

[24] F. Wu, F. Lyu, J. Ren, P. Yang, K. Qian, S. Gao, and Y. Zhang, “Characterizing internet card user portraits for efficient churn prediction model design,” IEEE Transactions on Mobile Computing, vol. 23, no. 2, pp. 1735–1752, 2024. https://doi.org/10.1109/TMC.2023.3241206.

[25] N. Ahmad, M. J. Awan, H. Nobanee, A. M. Zain, A. Naseem, and A. Mahmoud, “Customer personality analysis for churn prediction using hybrid ensemble models and class balancing techniques,” IEEE Access, vol. 12, pp. 1865–1879, 2024. https://doi.org/10.1109/ACCESS.2023.3334641.

[26] Y.-J. Han, J. Moon, and J. Woo, “Prediction of churning game users based on social activity and churn graph neural networks,” IEEE Access, vol. 12, pp. 101 971–101 984, 2024. https://doi.org/10.1109/ACCESS.2024.3429559.

[27] U. Gani Joy, K. E. Hoque, M. Nazim Uddin, L. Chowdhury, and S.-B. Park, "A big data-driven hybrid model for enhancing streaming service customer retention through churn prediction integrated with explainable AI," IEEE Access, vol. 12, pp. 69 130–69 150, 2024. https://doi.org/10.1109/ACCESS.2024.3401247.

[28] M. K. Rongala, "Comparative evaluation of machine learning models for customer churn prediction in the telecom sector," in 2025 International Conference on Computing Technologies (ICOCT), 2025, pp. 1–6. https://doi.org/10.1109/ICOCT64433.2025.11118675.

[29] A. Pravin, B. L. S. Bizotto, M. Sathiyanarayanan, and T. Jacob, "Enhanced framework to predict customer churn using machine learning," in 2025 International Conference on Inventive Computation Technologies (ICICT), 2025, pp. 8–12. https://doi.org/10.1109/ICICT64420.2025.11005219.

[30] S. A. Reddy, M. Gowtham, K. P. Tripathy, and M.Srinivas, "Enhanced telecom customer churn prediction using machine learning and deep learning models," in 2025 International Conference on Artificial Intelligence and Data Engineering (AIDE), 2025, pp. 280–285. https://doi.org/10.1109/AIDE64228.2025.10987431.