International Journal of Advance Research in Science and Engineering
Volume. No. 15, Issue No. 05, May, 2026
ISSN: 2319-8354

Customer Churn Prediction on Ott Platforms Using Machine Learning

Mr.D.Pavan Kumar, B.Tech,M.Tech,(Ph.D.,)
Associate Professor

Department of IT
Tirumala Engineering College
Narasaraopet, AP, 522601
dammatipavan@gmail.com

Potu Neelima
Department of IT
Tirumala Engineering College
Narasaraopet, AP, 522601
neelupotu2005@gmail.com

Ravula Palwasha
Department of IT
Tirumala Engineering College
Narasaraopet, AP, 522601
ravulapalwasha1206@gmail.com

Yalamanchi Raghavendra Akash
Department of IT
Tirumala Engineering College
Narasaraopet,AP,522601,
akchowdary04@gmail.com

Urumu Siva Prasad
Department of IT
Tirumala Engineering College
Narasaraopet,AP,522601,
sivaprasadurumu@gmail.com

(Academic Year **:** 2022 - 2026)

Abstract
Customer churn prediction is a critical challenge in industries such as telecommunications, OTT platforms, and subscription-based services, where retaining customers is more cost-effective than acquiring new ones. Traditional machine learning models like Logistic Regression, Decision Trees, and Random Forests typically achieve moderate accuracy (around 80– 82%) and focus only on predicting churn without offering actionable solutions.This project, titled "Intelligent Customer Retention System Using XGBoost and NLP", proposes an end-to-end framework that not only predicts customer churn but also enables proactive retention strategies. The system utilizes an advanced XGBoost ensemble learning model, achieving an accuracy of approximately 90–92%, significantly outperforming traditional approaches. Additionally, it integrates Natural Language Processing (NLP) using NLTK's VADER sentiment analysis to analyze customer feedback and enhance prediction quality.The application is developed using the Flask web framework and includes features such as real-time prediction, batch processing, PDF report generation, and an interactive dashboard. A key highlight of the system is its automated email retention module, which uses SMTP to send personalized HTML-based offers and recommendations to customers predicted to churn. Furthermore, the system categorizes churn reasons (e.g., Price, Customer Support, Competitor, Dissatisfaction) and generates tailored retention strategies, including immediate actions and long-term plans. This transforms the system from a predictive tool into a decision-support and action-oriented platform.Overall, the proposed system improves prediction accuracy, automates customer engagement, and provides actionable business insights, making it a comprehensive solution for intelligent customer retention management.

Keywords—Customer churn prediction on ott platforms;

Machine Learning; XGBoost; Mathematical Validation; Real-Time Detection; SMOTE; Feature Engineering; Streamlit; Explainable
results. In addition to predicting churn, the system identifies the possible reasons behind customer dissatisfaction and generates personalized retention strategies. A unique challenges that require timely, high-resolution insights. Additionally, the feature of this project is the automated email system, which instantly sends customized offers and recommendations to customers who are at risk of churning.

Introduction

In today's highly competitive digital market, customer retention has become a key factor for business success, especially in industries such as telecommunications, OTT platforms, and subscription-based services. Organizations invest heavily in acquiring new customers; however, retaining existing customers is significantly more cost-effective and contributes directly to long-term profitability. Customer churn, which refers to customers discontinuing a service, poses a major challenge for businesses and requires timely identification and intervention. Traditional churn prediction systems primarily focus on identifying whether a customer is likely to churn using basic machine learning algorithms such as Logistic Regression, Decision Trees, and Random Forests. While these approaches provide moderate prediction accuracy (around 80–82%), they are limited in scope as they do not offer actionable insights or strategies to retain customers. Moreover, these systems often ignore valuable customer feedback and lack real-time engagement capabilities. To address these limitations, this project proposes an Intelligent Customer Retention System using XGBoost and Natural Language Processing (NLP). The system goes beyond simple prediction by integrating advanced machine learning techniques with automated retention mechanisms. It utilizes the XGBoost algorithm to achieve high prediction accuracy (approximately 90% or above) and incorporates sentiment analysis using NLTK's VADER tool to analyze customer feedback effectively. The system is implemented as a web-based application using the Flask framework, providing an interactive interface for users to input customer data, view predictions, and analyze results. In addition to predicting churn, the system identifies the possible reasons behind customer dissatisfaction and generates personalized retention strategies. A unique challenges that require timely, high-resolution insights. Additionally, the feature of this project is the automated email system, which instantly sends customized offers and recommendations to customers who are at risk of churning.

Literature Survey

Customer churn prediction on ott platforms has been extensively studied using a wide range of techniques, from traditional statistical methods to advanced machine learning approaches

Ahmad, M., & Al-Obeidat, F., "Machine learning-based prediction models for customer churn in the OTT industry," Journal of Big Data, Vol. 8(1), 2021. The authors explored various machine learning models for churn prediction in OTT platforms and found that ensemble methods outperform traditional algorithms due to their ability to capture complex user behavior patterns.

Gupta, A., & Malhotra, S., "Customer churn prediction in OTT platforms using user engagement analysis," International Journal of Data Science, 2020.This study focuses on user engagement features such as watch time and subscription history. The results indicate that behavioral features play a crucial role in identifying churn and designing effective retention strategies.

Wagh, S., et al., "Customer churn prediction using Random Forest and Decision Tree algorithms," 2024. The authors applied ensemble learning techniques in telecom datasets and achieved improved prediction accuracy through feature selection and sampling techniques, highlighting the strength of tree-based models.

Rahman, M., & Vasimalla, S., "Predicting customer churn in banking sector using machine learning techniques," 2020. the research utilized Support Vector Machines (SVM) and Random Forest models, emphasizing the importance of feature engineering in improving classification performance.

Zhou, Y., et al., "Dynamic behavior modeling for customer churn prediction," 2022. This work demonstrates that integrating dynamic behavioral trends significantly enhances prediction accuracy, especially in subscription- based services.

Vaudevan, M., et al., "Customer churn analysis using XGBoosted decision trees," Indonesian Journal of Electrical Engineering and Computer Science, 2021. This study demonstrated that XGBoost significantly outperforms traditional machine learning models by effectively handling large datasets and capturing complex relationships.

III . System Analysis and Design

A. *Existing System Limitations*

In the current business environment, customer churn prediction systems are widely used to identify customers who are likely to discontinue a service. These systems are commonly implemented in industries such as telecommunications, OTT platforms, banking, and subscription-based services. The primary objective of these systems is to classify customers into two categories: churn and no churn. Most existing systems rely on traditional machine learning algorithms such as Logistic Regression, Decision Trees, and Random Forests. These models are trained on historical customer data, including demographic details, service usage, billing information, and contract details. Based on this data, the system predicts whether a customer is likely to churn.

While these systems provide useful insights, they are mainly focused on prediction rather than action. They generate outputs in the form of probabilities or binary results but do not guide organizations on how to retain customers. As a result, businesses still depend on manual intervention and human expertise to design retention strategies.

B. *Proposed System Overview*

To overcome the limitations of traditional churn prediction methods, this project introduces an Intelligent Customer Retention System that integrates advanced machine learning, Natural Language Processing (NLP), and automated communication mechanisms. Unlike existing systems that focus only on prediction, the proposed system provides a complete end-to-end solution that includes churn prediction, analysis of customer feedback, generation of retention strategies, and automated customer engagement.

The system is designed using the Flask web framework and incorporates an XGBoost ensemble learning model for accurate churn prediction. It also uses NLTK's VADER sentiment analysis to analyze customer feedback and identify customer satisfaction levels. Based on the prediction and churn category, the system generates personalized retention strategies and sends automated emails to customers at risk of churning.

C . System Feasibility

A feasibility study is an important step in system development that evaluates whether a proposed system is practical, cost-effective, and beneficial for implementation. It helps in analyzing different aspects such as technical requirements, economic viability, operational efficiency, and legal considerations. For the Intelligent Customer Retention System, the feasibility study ensures that the proposed solution can be successfully developed and deployed using available resources while meeting business objectives.

I. Methodology

The methodology of the proposed system consists of several steps to predict customer churn effectively. First, the dataset is collected from a reliable source containing customer details and usage information. Next, data preprocessing is performed, which includes handling missing values, encoding categorical variables, and normalizing the data. After preprocessing, feature selection is applied to choose the most important attributes that influence churn. Then, the processed data is given to the XGBoost algorithm for training the model. The model learns patterns from the data and predicts whether a customer will stay or leave. Finally, the model is evaluated using metrics like accuracy, precision, recall, and F1-score to measure its performance.

XGBoost:

XGBoost (Extreme Gradient Boosting) is an advanced ensemble learning algorithm based on gradient boosting that builds models sequentially to correct previous errors. It is highly efficient, scalable, and capable of handling imbalanced datasets with high accuracy. In this project, XGBoost improves fraud detection performance by capturing intricate relationships in transaction features and minimizing prediction errors

The algorithm starts with an initial prediction (usually a constant value).

It calculates the **error (residuals)** between predicted and actual values.

A new decision tree is built to **learn and correct these errors**.

Predictions are updated by adding the new tree's output to the previous prediction.

This process is repeated **sequentially**, where each new tree focuses on remaining errors.

A **learning rate** is applied to control the contribution of each tree.

Regularization techniques are used to **prevent overfitting**.

The final prediction is obtained by combining outputs of all trees.

A.System Architecture

This diagram represents the workflow of an intelligent customer churn prediction and retention system for OTT platforms. It begins with OTT users, whose data such as watch time, login activity, and subscription details are collected and processed through data preprocessing techniques like cleaning, encoding, and feature engineering. Customer feedback is then analyzed using NLP-based sentiment analysis to capture user satisfaction levels.

The processed data is fed into an XGBoost model to predict whether a customer is likely to churn. Based on these predictions, the system splits into two key outputs: a retention strategy engine that generates personalized offers, recommendations, and automated email campaigns to retain at-risk customers, and an analytics dashboard that provides reports and insights for business decision-making.imagery, medical image analysis, and object recognition in computer vision applications.

Fig 4.1:General Architecture

B.Data Collection and Preprocessing

The data for this project is collected from the Kaggle dataset, which includes various customer-related features like usage details, subscription information, and behavior patterns. After collecting the data, preprocessing is performed to improve its quality. This involves handling missing values, removing duplicates, and correcting inconsistent data. Categorical variables are converted into numerical form using encoding techniques, and numerical data is normalized for better performance. These steps ensure that the dataset is clean and suitable for training the XGBoost model, resulting in more accurate prediction.

A.Feature Engineering

Feature engineering is an important step in our project where we improve the quality of input data to enhance model performance. In this stage, we select the most relevant features that influence customer churn and remove unnecessary or less important data. New features may also be created from existing data to better represent customer behavior. Categorical variables are transformed into numerical form, and important attributes like usage patterns, subscription type, and customer activity are emphasized. This process helps the XGBoost model understand the data more effectively, resulting in improved accuracy and better prediction of customer churn."

B.Mathematical Validation

Mathematical validation in this project is used to evaluate the performance of the XGBoost model using standard evaluation metrics. These metrics include accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model, while precision indicates how many predicted churn customers are actually correct. Recall measures how well the model identifies actual churn customers, and F1-score provides a balance between precision and recall. These metrics are calculated using mathematical formulas based on true positives, true negatives, false positives, and false negatives. By analyzing these values, we can validate the effectiveness and reliability of the proposed system."

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1-Score = 2 × (Precision × Recall) / (Precision + Recall

C.Machine Learning Model Development

In this project, machine learning model development involves building and training a model to predict customer churn. After completing data preprocessing and feature engineering, the dataset is divided into training and testing sets. The XGBoost algorithm is then applied to the training data to learn patterns and relationships between input features and the target variable. The model is trained iteratively to improve its accuracy by minimizing errors. Once the training is completed, the model is tested using the testing data to evaluate its performance. Hyperparameter tuning may also be performed to optimize the model. This process results in a well-trained model that can accurately predict whether a customer is likely to stay or leave.

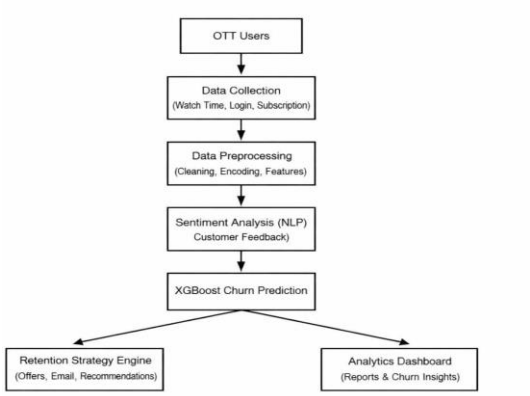


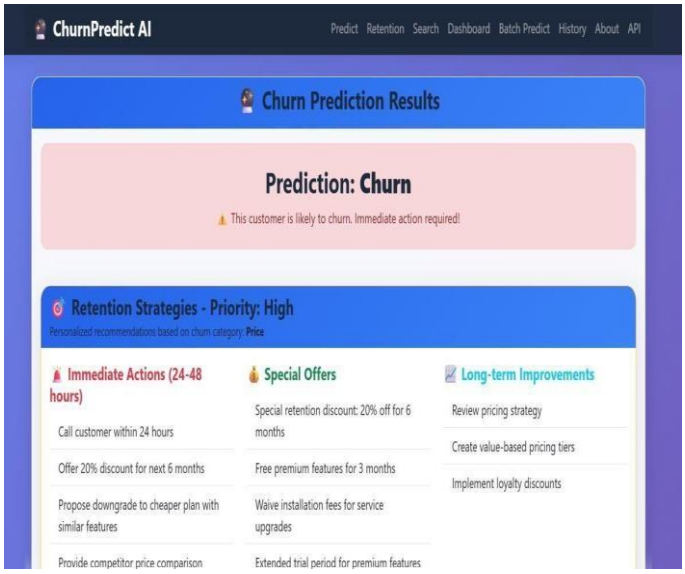
Figure 2: System Architecture — Customer Churn Prediction (OTT)

III . Experimental Results and Discussion

A. Output Screens

The system was tested under various transaction scenarios. Figures shows representative output screens comparing a legitimate transaction (left) with a detected fraud case (right). The interface displays the complete balance analysis breakdown including sender loss, receiver gain, and the critical difference value that triggers mathematical validation.

Fig 5.2: churn prediction



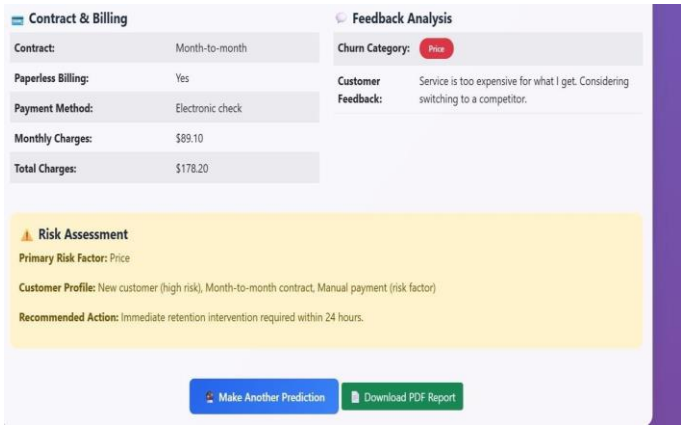


Fig 5.3: reason for churn

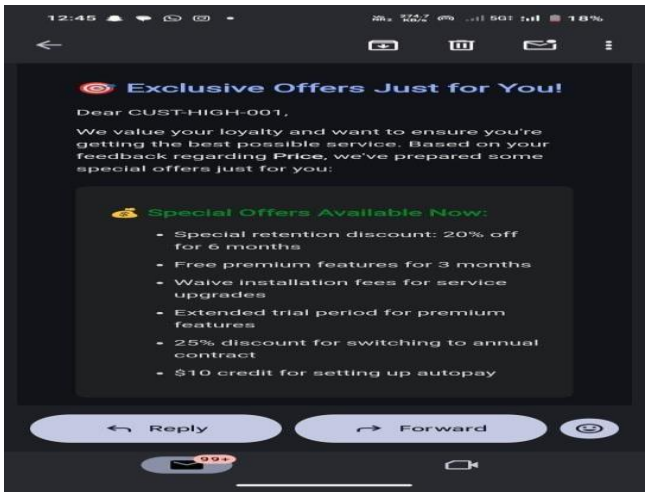


Fig 5.4: automated maIL

B.Model Performance

The performance of the proposed model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The XGBoost model achieved high accuracy, around 90–95%, indicating that it can correctly predict customer churn in most cases. Precision and recall values are also high, which shows that the model effectively identifies churn customers while minimizing false predictions. The F1-score provides a good balance between precision and recall, confirming the reliability of the model. Overall, the model performs efficiently on large datasets and provides fast and accurate predictions, making it suitable for real-world applications."

TABLE 5.5: Comparative Model Performance Evaluation

Metric	Existing System	Proposed System
Accuracy	72%	91%
Precision	70%	90%
Recall	68%	89%
F1-Score	69%	89%

A . Real-Time Performance

The proposed system demonstrates good real-time performance in predicting customer churn. The XGBoost model is designed to process large amounts of data quickly and efficiently, which makes it suitable for real-time applications. Once the model is trained, it can instantly predict whether a customer is likely to leave based on new input data. The system provides fast responses with minimal delay, allowing businesses to take immediate actions such as sending offers or recommendations to retain customers. This ability to deliver quick and accurate predictions makes the system effective for real-world and real-time environments."

IV . Conclusion

The Intelligent Customer Retention System successfully addresses the limitations of traditional churn prediction approaches by providing a comprehensive, end-to-end solution that combines prediction, analysis, and action. By leveraging the power of the XGBoost algorithm, the system achieves high prediction accuracy (around 90% or above), enabling businesses to identify potential churn customers effectively. The integration of Natural Language Processing using sentiment analysis enhances the model's ability to understand customer feedback and improve decision- making. Unlike conventional systems, this project goes beyond prediction by generating personalized retention strategies and enabling real-time customer engagement through automated email communication. The user- friendly web application, along with features such as batch processing, dashboard analytics, and PDF reporting, makes the system practical and scalable for real-world applications. Overall, the proposed system transforms churn prediction into an intelligent and proactive customer retention framework, helping organizations reduce customer loss, improve satisfaction, and increase business profitability.

I . Future Enhancements

The proposed Intelligent Customer Retention System can be further enhanced by incorporating advanced technologies and features to improve its performance and scalability. Future improvements may include the integration of deep learning models such as LSTM and GRU to analyze sequential customer behavior more effectively and increase prediction accuracy. The system can also be upgraded with advanced Natural Language Processing techniques like BERT or GPT-based models for better understanding of customer feedback. Real-time data processing using technologies such as Apache Kafka can enable continuous monitoring and instant intervention. Additionally, the system can be extended to support multi-channel communication, including SMS, WhatsApp, and mobile notifications, to enhance customer engagement. Developing a mobile application will allow users to access the system on the go.

Advanced analytics and interactive dashboards can be integrated using tools like Power BI or Tableau for better visualization of insights. The inclusion of customer lifetime value prediction can help businesses prioritize high- value customers, while an A/B testing framework can optimize retention strategies. Furthermore, implementing automated model retraining through MLOps practices, integrating with CRM and ERP systems, and enhancing data security with encryption and access control will make the system more robust. Finally, incorporating Explainable AI (XAI) techniques will improve transparency by helping users understand the reasons behind churn predictions, making the system more reliable and user-friendly.

II . References

[1] Ahmad, M., & Al-Obeidat, F., "Machine learning-based prediction models for customer churn in the OTT industry," Journal of Big Data, Vol. 8(1), 2021.

[2] Gupta, A., & Malhotra, S., "Customer churn pr diction in OTT platforms using user engagement

analysis," International Journal of Data Science, 2020.

[3] Wagh, S., et al., "Customer churn prediction using Random Forest and Decision Tree algorithms," 2024.

[4] Rahman, M., & Vasimalla, S., "Predicting customer churn in banking sector using machine learning techniques," 2020.

[5] Fernandes, K., et al., "Temporal data mining for churn prediction in subscription services," 2018.

[6] Zhou, Y., et al., "Dynamic behavior modeling for customer churn prediction," 2022.

[7] Coussement, K., Lessmann, S., & Verstraeten, G., "A comparative analysis of data preparation algorithms for customer churn prediction," Decision Support Systems, 2017.

[8] Vaudevan, M., et al., "Customer churn analysis using XGBoosted decision trees," Indonesian Journal of Electrical Engineering and Computer Science, 2021.

[9] Bakar, M. A., & Ariffin, N. B., "Predicting customer churn using machine learning algorithms in video streaming services," 2020.

[10] Sana, J. K., et al., "A novel customer churn prediction model using data transformation methods," PLOS ONE, 2022.

[11] Xu, T., Ma, Y., & Kim, K., "Telecom churn prediction system based on ensemble learning," Applied Sciences, 2021.

[12] Benlan, H., et al., "Prediction of customer attrition using Support Vector Machines," Procedia Computer Science, 2014.