IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁘ Peer-reviewed & Refereed journal ⁘ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15820

Explainable AI-Based Customer Churn Prediction Using Machine Learning and SHAP Analysis

Dr. Bharathi M P[1], Narendra C[2], Abhishekaraddi Maraddi [3]

Assistant Professor, Department of Master of Computer Applications, Surana College (Autonomous), Bengaluru, India[1]

Student, Master of Computer Applications, Surana College (Autonomous), Bengaluru, India[2]

Student, Master of Computer Applications, Surana College (Autonomous), Bengaluru, India[3]

Abstract: Over Customer churn prediction plays a critical role in withstanding the competition in the contemporary business world. This paper presents an Explainable Artificial Intelligence (XAI) framework for predicting customer churn using advanced machine learning algorithms. The framework employed includes XGBoost, LightGBM, Random Forest, and Ensemble Learning to develop an accurate predictive model. The study used customer demographics and behavior data to train the algorithm. Additionally, the study utilized SHAP (SHapley Additive ExPlanations) to interpret the results by illustrating the key factors associated with customer churn. The results indicated that all the predictive models had good Accuracy, F1-score, and ROC-AUC performance metrics. However, the Ensemble Learning model had the highest ROC-AUC of 93.8%. Thus, the study concludes that the proposed solution is an effective approach to predicting customer churn and aids businesses in making the right decisions based on the results.

Keywords: Explainable AI, SHAP, XGBoost, LightGBM, Random Forest, Ensemble Learning, Machine Learning, Data Analytics.

I INTRODUCTION

Customer churn is a critical business challenge that results in significant revenue loss. Traditional churn detection methods are often inefficient, while machine learning and analytics improve prediction accuracy by identifying customer behavior patterns. However, cloud-based prediction systems demand high computational resources, increasing processing time and operational costs. Even a 5% reduction in churn can boost profits by 25–95%, while winning back lost customers costs 3–7 times more than acquiring new ones. This underscores the need for lightweight, efficient, and explainable churn prediction models that can be deployed in real-world business environments without heavy reliance on cloud infrastructure.

Machine learning enables systems to learn patterns from data and make accurate predictions without explicit programming, with ensemble methods like XGBoost, LightGBM, and Random Forest being especially effective for structured, tabular problems such as churn prediction. However, since these models often act as "black boxes," explainable AI techniques like SHAP are increasingly used to interpret predictions and build trust in real-world decision-making.

Explainable Artificial Intelligence (XAI) can help organizations gain insight and trust in Machine Learning decisions by making it transparent to the users regarding the outcomes of the predictions. This system can make accurate churn forecasts, yet remain interpretable and relevant to business. Customer demographic and behavioral data is analysed using Machine Learning Algorithms like XGBoost, LightGBM, Random Forest and Ensemble Learning customer's churn. The research employs sophisticated machine learning models and techniques for model interpretability to develop an Explainable AI system for customer churn prediction, which provides accurate and transparent churn forecasts.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026
DOI: 10.17148/IJARCCE.2026.15820

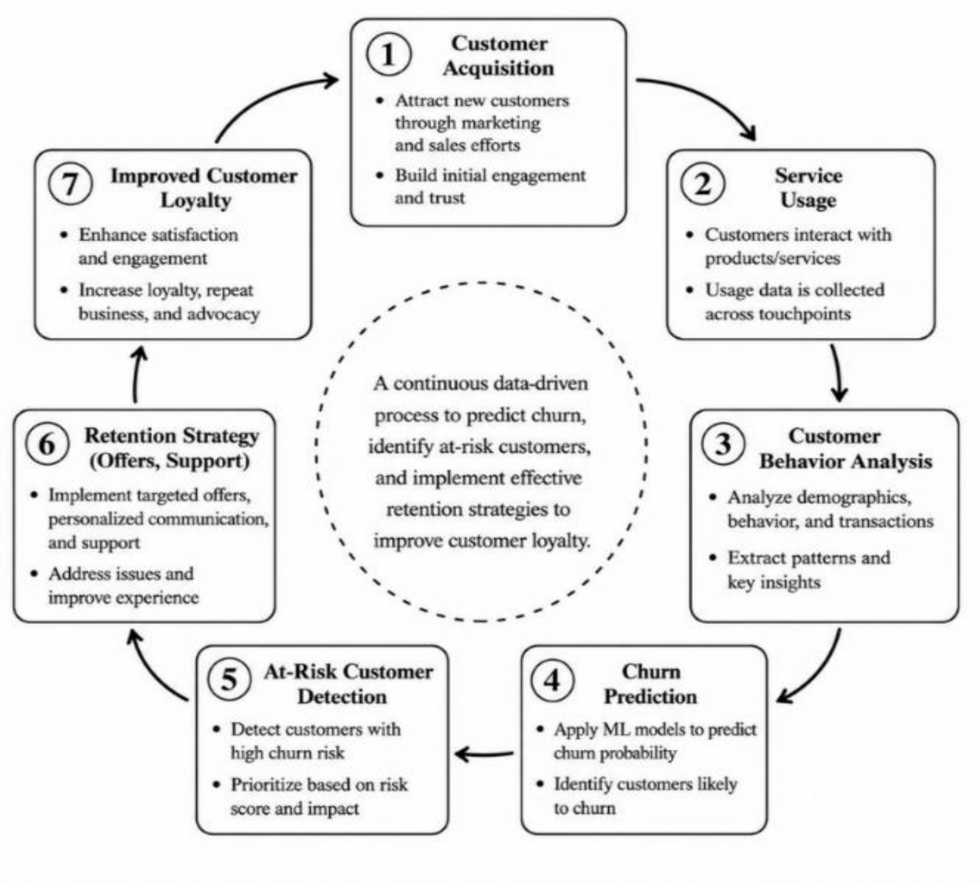


Figure 1: Customer Retention Lifecycle

Figure 1 illustrates the Customer Retention Lifecycle, which begins with customer acquisition and service usage, followed by the analysis of demographic, behavioral, and transactional data to identify customer behavior patterns. Machine learning models are then used to predict churn and detect at-risk customers based on churn probability. Personalized retention strategies, such as targeted offers, loyalty programs, and proactive customer support, are implemented to improve customer satisfaction and loyalty. This continuous data-driven process helps organizations reduce churn, strengthen customer relationships, and achieve sustainable business growth.

II. LITERATURE REVIEW

Customer churn prediction has gained significant attention in recent years due to its importance in customer retention and business profitability. Several researchers have applied machine learning techniques to identify customers who are likely to discontinue services. Traditional algorithms such as Logistic Regression, Decision Trees, and Support Vector Machines have been widely used for churn prediction and have shown satisfactory performance across various domains.

A study [1] introduced the Random Forest algorithm, an ensemble learning technique that combines multiple decision trees using bootstrap aggregation and majority voting. The algorithm effectively reduces overfitting while improving prediction accuracy and generalization. Due to its ability to handle high-dimensional and noisy datasets, Random Forest has become one of the most widely used machines learning algorithms for customer churn prediction. XGBoost, an optimized gradient boosting algorithm designed to improve computational efficiency and predictive performance. By incorporating regularization, parallel processing, and tree pruning techniques, XGBoost achieves superior accuracy while reducing overfitting. Several studies have demonstrated that XGBoost consistently outperforms traditional machine learning algorithms in customer churn prediction tasks [2].

LightGBM is an efficient gradient boosting framework that employs histogram-based learning and leaf-wise tree growth to achieve fast training, low memory usage, and high predictive accuracy, making it well-suited for large-scale customer churn prediction [3]. A group of researchers introduced SHAP (SHapley Additive Explanations), an Explainable Artificial Intelligence framework based on cooperative game theory. SHAP assigns contribution values to each feature involved in

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁚ Peer-reviewed & Refereed journal ⁚ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15820

a prediction and provides both global and local explanations. Its theoretical consistency and compatibility with tree-based models have made SHAP one of the most widely adopted explainability techniques in customer churn prediction [4].

Ribeiro, Singh, and Guestrin [5] introduced LIME, a model-agnostic explainability framework that provides local explanations for individual predictions, improving transparency and user trust in machine learning models. A study combined Topological Data Analysis with machine learning using the IBM Telco Customer Churn dataset (7,043 records, 21 features). XGBoost achieved the highest performance with 98.5% accuracy, demonstrating the effectiveness of topology-based feature extraction [6].

Idris, Khan, and Lee [7] combined machine learning with social network analysis to improve telecom customer churn prediction by incorporating customer interaction patterns alongside demographic and service usage data which obtained 93% of AUC value. A telecom customer churn dataset was utilized (3,150 records, 14 attributes) to compare ML and deep learning models. ANN achieved the best performance with 99% accuracy, outperforming CNN and conventional machine learning algorithms [8].

Ke Peng et al. (2023) proposed a GA-XGBoost model using the BankChurners.csv dataset (10,127 records, 19 features). The model achieved 96.27% accuracy and 0.97 AUC, improving prediction accuracy and interpretability through genetic optimization. [9]. Recent studies integrated SHAP with churn prediction models to explain key factors influencing customer churn while maintaining high predictive performance and supporting informed business decisions. Adaptive ensemble models combining Random Forest, XGBoost, and LightGBM achieved higher accuracy, precision, recall, and robustness than individual classifiers for customer churn prediction [10, 11].

[12] Zhang and Liu combined Genetic Algorithms, XGBoost and SHAP to optimize feature selection with improved churn prediction accuracy of and provide interpretable feature importance. Studies have demonstrated that deep learning effectively predicts telecom customer churn but highlighted the need for explainability to enhance model transparency and the ensemble methods, particularly Random Forest and XGBoost, outperformed traditional machine learning models in telecom churn prediction [13,14].

[15] Ibrahim, Ahmed, and Hassan proposed an explainable ensemble framework using SHAP to improve churn prediction and identify key factors affecting customer attrition in the energy sector. Authors [16, 17] showed that integrating Explainable AI with machine learning, ensemble, and deep learning models improves both churn prediction and business decision-making.

III PROPOSED RESEARCH

A. MOTIVATION AND PRELIMINARIES OF WORK

The proposed research presents an Explainable AI-based Hybrid Customer Churn Prediction Framework that integrates XGBoost, LightGBM, Random Forest, and Ensemble Learning to improve prediction accuracy and robustness. The framework employs SHAP to explain model predictions and identify the most influential factors contributing to customer churn. By combining high predictive performance with interpretability, the proposed system enables proactive customer retention and supports reliable, data-driven business decisions across various industries.

B. OBJECTIVES

The proposed research aims to:

1. Develop an accurate and scalable customer churn prediction framework by integrating XGBoost, LightGBM, Random Forest, and Hybrid Ensemble Learning techniques.
2. Enhance the transparency and interpretability of churn predictions using SHAP-based Explainable Artificial Intelligence (XAI) to identify the key factors influencing customer churn.
3. Support proactive customer retention by accurately identifying at-risk customers and providing explainable insights for informed business decision-making across multiple domains.

C. DATASET DESCRIPTION

The study utilizes a Customer Churn dataset containing demographic, behavioral, and service-related attributes such as Age, Gender, Tenure, Usage Frequency, Contract Length, Subscription Type, Payment Delay, and Total Spend to predict customer churn. Data preprocessing involved removing irrelevant attributes, encoding categorical features, and splitting the dataset into 70% training and 30% testing sets. The processed data was used to train XGBoost, LightGBM, Random

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15820

Forest, and Ensemble Learning models. Furthermore, SHAP-based Explainable AI (XAI) was employed to interpret model predictions and identify key factors influencing customer churn, enabling effective retention strategies.

D. METHODOLOGY

The proposed framework follows a systematic approach involving data preprocessing, model training, evaluation, and explainability analysis. After preprocessing, the prepared dataset is used for model development and validation to ensure reliable predictive performance.

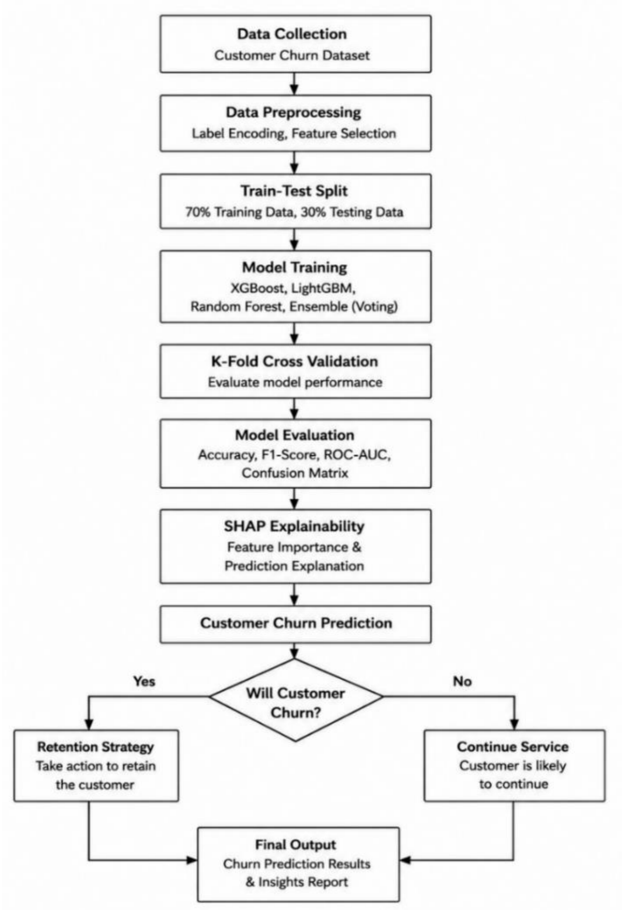


Figure 2: Workflow of the proposed model

The above Figure 2 presents the proposed customer churn prediction workflow, which begins with customer data collection, followed by preprocessing using feature selection and label encoding. The processed dataset is split into 70% training and 30% testing sets, and XGBoost, LightGBM, Random Forest, and Ensemble Learning models are trained and validated using K-Fold Cross Validation with hyperparameter tuning. Model performance is evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and Confusion Matrix. SHAP-based Explainable AI (XAI) is then applied to interpret predictions and identify key factors influencing churn. Based on the predicted outcome, at-risk customers receive personalized retention strategies, while the framework generates actionable churn prediction results to support informed business decision-making

XGBoost, LightGBM, Random Forest, and Ensemble Learning models are trained using hyperparameter tuning and 5-fold cross-validation. The models are evaluated using Accuracy, F1-Score, ROC-AUC, and Confusion Matrix. Finally, SHAP-based Explainable AI (XAI) is employed to interpret predictions, identify key churn factors, and support effective customer retention strategies.

The customer churn dataset was preprocessed using label encoding for categorical features (Gender, Subscription Type, and Contract Length) and normalization for numerical features (Age, Tenure, Usage Frequency, Support Calls, Payment Delay, Total Spend, and Last Interaction). This reduced scale variations, improved model training, and retained the target variable (Churn), where 1 = churned and 0 = retained.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15820

CustomerID	Age	Gender	Tenure	Usage Frequency	Support Calls	Payment Delay	Subscription Type	Contract Length	Total Spend	Last Interaction	Churn
1	-1.6387	0	-0.5131	-0.5841	-0.6536	1.1457	0	1	-0.2169	-1.7837	1
2	-0.1492	0	-0.3770	0.7881	0.2821	-0.4248	2	1	-0.2565	-0.0452	0
3	0.3620	1	-0.4228	-1.0271	-1.1210	1.2896	1	0	0.2455	0.1474	0
4	-0.5348	1	-1.2770	-0.8056	-0.1856	0.0977	1	2	-1.2982	-0.1473	0
5	0.9474	0	1.0320	0.3298	0.7226	-1.5037	2	0	-0.3900	-0.1473	0

Table 1: Data fields after pre-processing

The customer churn prediction system was developed by preprocessing the dataset, applying SMOTE to address class imbalance, and training XGBoost, Random Forest, LightGBM, and Gradient Boosting models. A Hybrid AI model was built using a soft voting ensemble of XGBoost, Random Forest, and LightGBM to improve prediction accuracy and robustness. The models were evaluated using cross-validation and performance metrics, while SHAP-based Explainable AI (XAI) was employed to interpret predictions and identify key factors influencing customer churn

V IMPLEMEMTATION

A. DATA PRE-PROCESSING

The proposed framework follows preprocessing the customer churn dataset to improve data quality and compatibility with machine learning algorithms. Irrelevant attributes are removed categorical features are encoded using Label Encoding, and feature selection is performed to retain significant attribute.

The Original dataset presents a sample customer churn dataset containing demographic, behavioral, financial, and service-related attributes used for churn prediction. Features such as Age, Gender, Tenure, Usage Frequency, Payment Delay, Total Spend, Subscription Type, Contract Length, and Last Interaction describe customer characteristics and behavior.

B. DATA FIELDS AFTER PRE-PROCESSING

The preprocessed customer churn data for machine learning is shown in the table below. Label encoding was applied for categorical variables like Gender, Subscription Type and Contract Length to convert them into numeric format. Numerical features were normalized so that they were on the same scale and have positive and negative values around zero, such as Age, Tenure, Usage Frequency, Support Calls, Payment Delay, Total Spend, and Last Interaction. This preprocessing helps to eliminate scale variations among features, which improves the performance which improves the performance training efficiency of models. The churned customer column (Churn) stays the same: 1 = churned customer, 0 = customer who is still using the service. The customer churn prediction system was developed by preprocessing the dataset, applying SMOTE to address class imbalance, and training XGBoost, Random Forest, LightGBM, and Gradient Boosting models. A Hybrid AI model was built using a soft voting ensemble of XGBoost, Random Forest.

Table 2: STEPS IN SHAP IMPLIMENTATION

STEP 1: Import SHAP Library
```
import shap
```
STEP 2: Create SHAP Explainer
```
explainer = shap.TreeExplainer(xgb_model)
```
STEP 3: Compute SHAP Values
```
shap_values =
explainer.shap_values(X_test)
```
STEP 4: Generate SHAP Summary Plot
```
shap.summary_plot(
    shap_values, X_test)
```
STEP 5: Generate SHAP Feature Importance Plot
```
shap.summary_plot(
shap_values,   X_test, plot_type="bar")
```

The above table 2 illustrates the SHAP-based Explainable AI (XAI) framework used to improve the transparency and interpretability of the customer churn prediction model. SHAP TreeExplainer computes feature contributions, while

IJARCCE
ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15820

summary and feature importance plots identify the key factors influencing churn predictions. This explainability enhances stakeholder trust and supports informed customer retention decisions.

SHAP (SHapley Additive exPlanations) is integrated into the proposed framework to improve the transparency and interpretability of customer churn predictions. It computes the contribution of each feature to an individual prediction based on Shapley values derived from cooperative game theory. The SHAP value for a feature is calculated as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!}\left[f(S \cup \{i\}) - f(S)\right]$$

where ϕ_i represents the contribution of feature i, N is the set of all features, S is a subset excluding feature i, and $f(\cdot)$ denotes the model prediction. SHAP identifies the most influential features affecting customer churn, enabling accurate interpretation of model decisions and supporting effective retention strategies.

VI. RESULTS AND DISCUSSION

The performance of all the implemented models was validated based on different performance metrics including Accuracy, F1-Score, ROC-AUC and Confusion Matrix. These metrics not only assist with model correctness but also in understanding the efficacy to distinguish between churn and non-churn customers. The results of the experiments showed that the highest accuracy was obtained by XGBoost and Ensemble Learning with 89.2%, followed by LightGBM with 89.1%, and Random Forest with 88.2%. The accuracy values are high, suggesting that the proposed framework can achieve high classification accuracy in predicting customer churn.

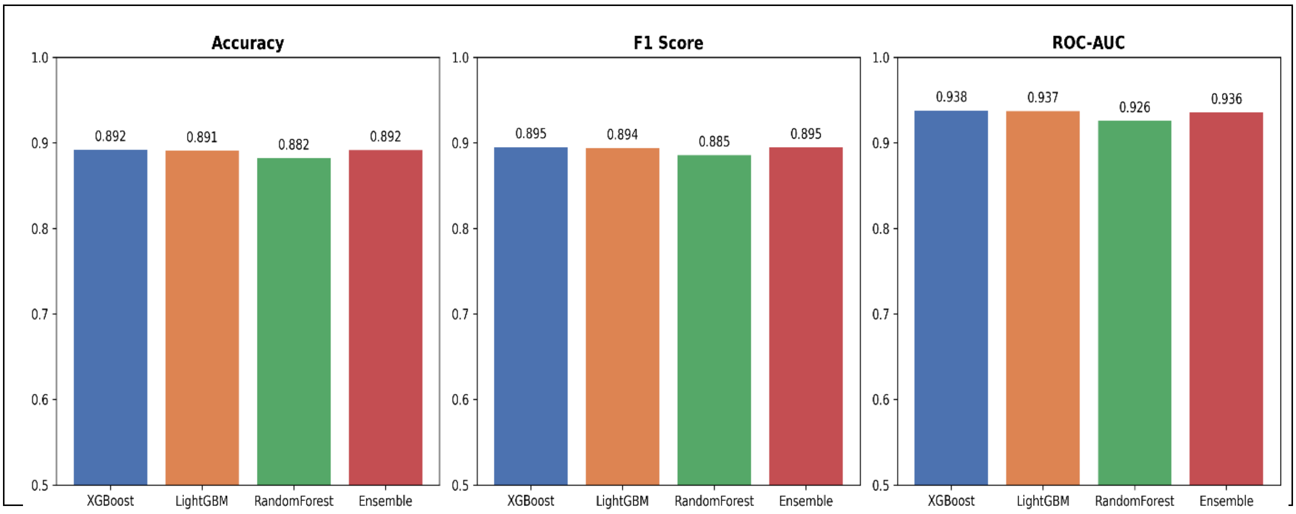


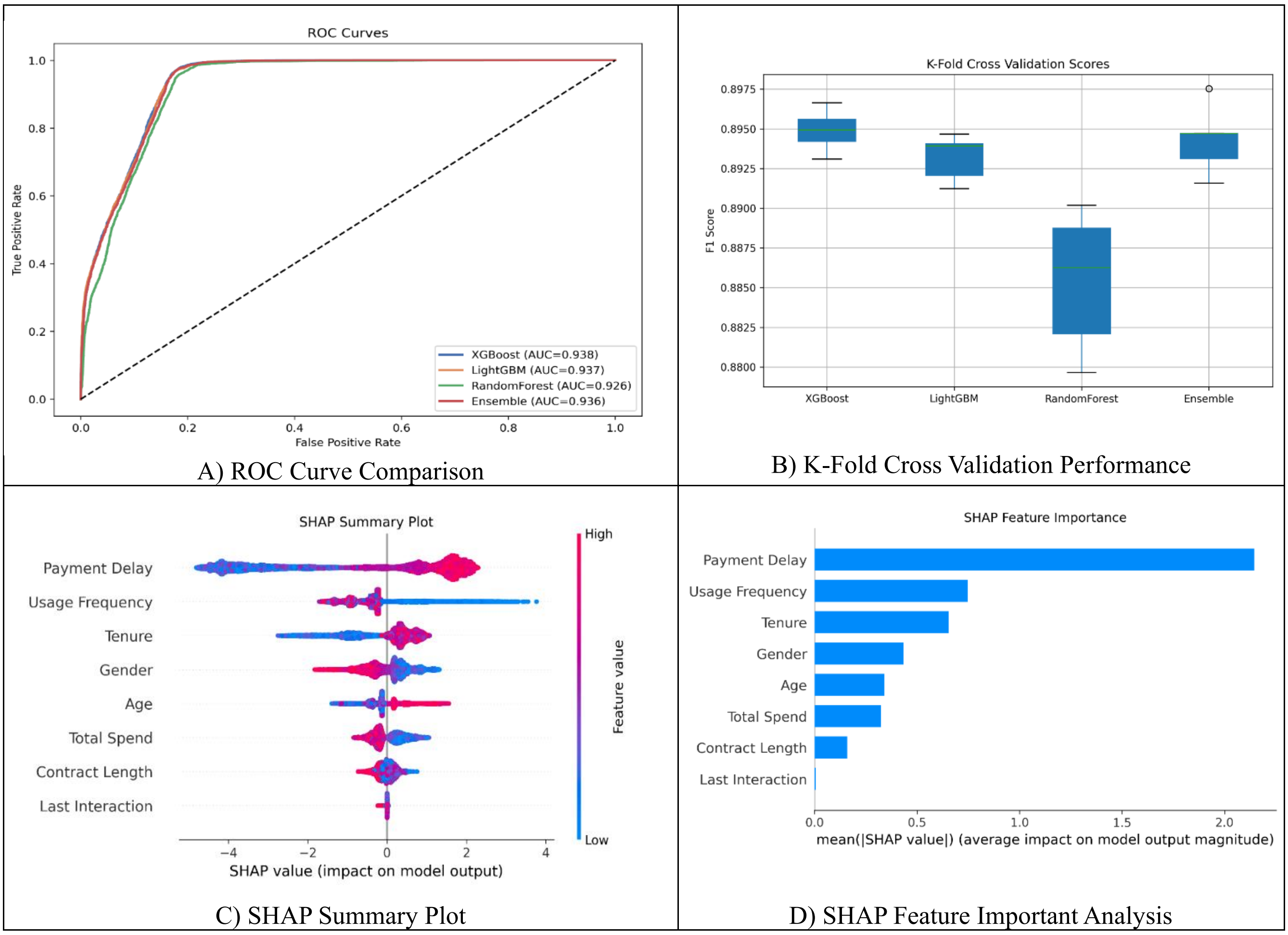
Figure 3 presents the comparative performance of XGBoost, LightGBM, Random Forest, and an Ensemble model across Accuracy, F1-Score, and ROC-AUC metrics for customer churn prediction. XGBoost achieved the best overall performance (Accuracy: 0.892, F1: 0.895, ROC-AUC: 0.938), narrowly ahead of LightGBM and the Ensemble, while Random Forest trailed slightly across all metrics. Despite these small variations, all four models demonstrated strong predictive effectiveness, with scores tightly clustered between 0.88 and 0.94. This consistency suggests that model choice for deployment could reasonably factor in efficiency and interpretability alongside raw performance.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15820



A) ROC Curve Comparison

B) K-Fold Cross Validation Performance

C) SHAP Summary Plot

D) SHAP Feature Important Analysis

Figure 4 Model Evaluation

Figure A presents the ROC curves of the four machine learning models, with XGBoost achieving the highest ROC-AUC score (0.938), indicating the best overall classification performance for customer churn prediction.

Figure B presents the K-Fold Cross-Validation F1-score distributions of the four machine learning models, showing that XGBoost achieved the most consistent and highest validation performance, while Random Forest exhibited greater variability across the folds.

Figure C presents the SHAP summary plot of the XGBoost model, indicating that Payment Delay, Usage Frequency, and Tenure are the most influential features affecting customer churn prediction, while Last Interaction has the least impact on the model's output.

Figure D presents the SHAP feature importance plot, showing that Payment Delay is the most influential feature in customer churn prediction, followed by Usage Frequency and Tenure, whereas Last Interaction has the least impact on the model's predictions.

DOI: 10.17148/IJARCCE.2026.15820

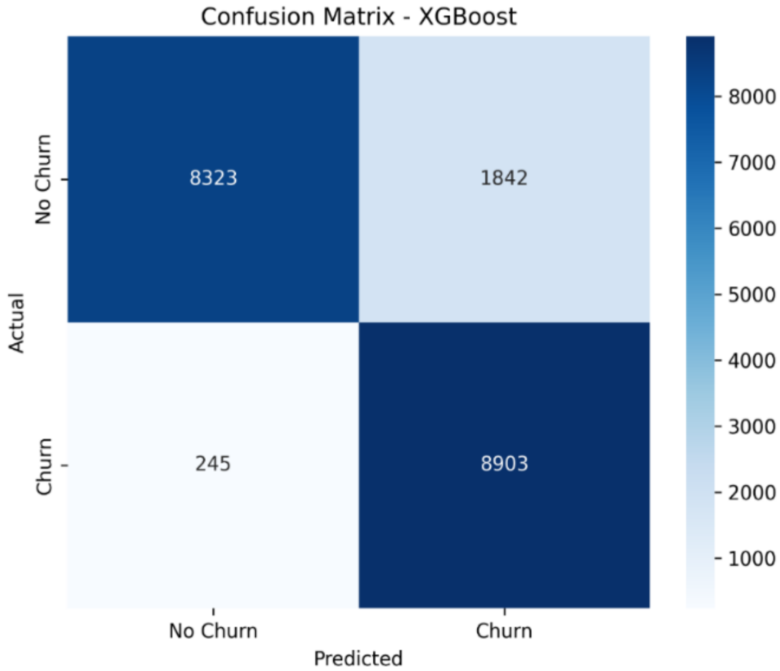


Figure 5: Confusion Matrix – XGBoost

Figure 5 presents the confusion matrix of the XGBoost model, showing 8,323 correctly classified non-churn and 8,903 correctly classified churn instances, with only 1,842 false positives and 245 false negatives. The notably low false negative count indicates that the model successfully identifies most at-risk customers, which is critical for effective retention intervention. Overall, the results confirm XGBoost's strong and reliable predictive performance for churn classification.

VII CONCLUSION

This research developed an Explainable Artificial Intelligence (XAI)-based customer churn prediction system using XGBoost, LightGBM, Random Forest, and Ensemble Learning models. After data preprocessing and model evaluation, XGBoost and Ensemble Learning achieved the best predictive performance in terms of Accuracy, F1-Score, and ROC-AUC. The integration of SHAP improved model interpretability by identifying the key factors influencing customer churn, such as Payment Delay, Usage Frequency, and Tenure. Overall, the proposed approach provides accurate and explainable churn predictions, enabling organizations to identify at-risk customers and implement effective retention strategies. Future work can focus on integrating deep learning techniques, larger real-world datasets, and real-time analytics to further enhance prediction performance and scalability.

REFERENCES

[1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, pp. 785–794, 2016.
[3] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 3146–3154, 2017.
[4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.
[5] M. T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 1135–1144, 2016.
[6] M. H. B. M. B. Ahmed, A. El Mhamdi, M. Albugami and A. Barnawi, "Using Topological Data Analysis and Machine Learning to Predict Customer Churn," *Journal of Big Data*, vol. 11, no. 1, 2024

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⫘ Peer-reviewed & Refereed journal ⫘ Vol. 15, Issue 8, August 2026

DOI: 10.17148/IJARCCE.2026.15820

[7] M. H. M. A. Idris, A. Khan and Y. S. Lee, "Customer Churn Prediction in Telecom Using Machine Learning and Social Network Analysis in Big Data Platform," *arXiv preprint*, arXiv:1904.00690, 2019.

[8] L. Saha, H. K. Tripathy, T. Gaber, H. El-Gohary and E.-S. M. El-Kenawy, "Deep Churn Prediction Method for Telecommunication Industry," *Sustainability*, vol. 15, no. 5, Art. no. 4543, 2023

[9] Peng K, Peng Y, Li W (2023) Research on customer churn prediction and model interpretability analysis. PLoS ONE 18(12): e0289724. https://doi.org/10.1371/journal.pone.0289724

[10] A. Almansoori et al., "Customer Churn Prediction Model Using Explainable Machine Learning," *arXiv preprint*, arXiv:2303.00960, 2023.

[11] A. Kumar, R. Sharma and S. Patel, "Enhancing Customer Churn Prediction in Telecommunications: An Adaptive Ensemble Learning Approach," *arXiv preprint*, arXiv:2408.16284, 2024.

[12] Y. Zhang and X. Liu, "Research on Telecom Customer Churn Prediction Based on GA-XGBoost and SHAP," *Journal of Computer and Communications*, vol. 10, no. 11, pp. 128–144, 2022.

[13] H. Almansoori, M. Almazrouei and A. Alkaabi, "Deep Churn Prediction Method for Telecommunication Industry," *Sustainability*, vol. 15, no. 5, 2023.

[14] M. A. Khan, S. Ahmed and M. Hussain, "Improved Decision Tree, Random Forest, and XGBoost Algorithms for Predicting Client Churn in the Telecommunications Industry," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 15, no. 12, 2024.

[15] S. Ibrahim, M. Ahmed and R. Hassan, "Explainable AI for Customer Churn Prediction in the Energy Sector Using Ensemble Machine Learning Models," *EAI Endorsed Transactions on Artificial Intelligence*, vol. 2, no. 1, 2025.

[16] A. Baby, P. Raj and S. Thomas, "Explainable Machine Learning for Bank Customer Churn Prediction: A Comparative Study of Traditional, Ensemble, and Deep Learning Models," 2025.

[17] Y. Wang, L. Chen and H. Zhao, "Bank Customer Churn Prediction Based on Stacking Model," *Advances in Engineering and Management*, 2025.