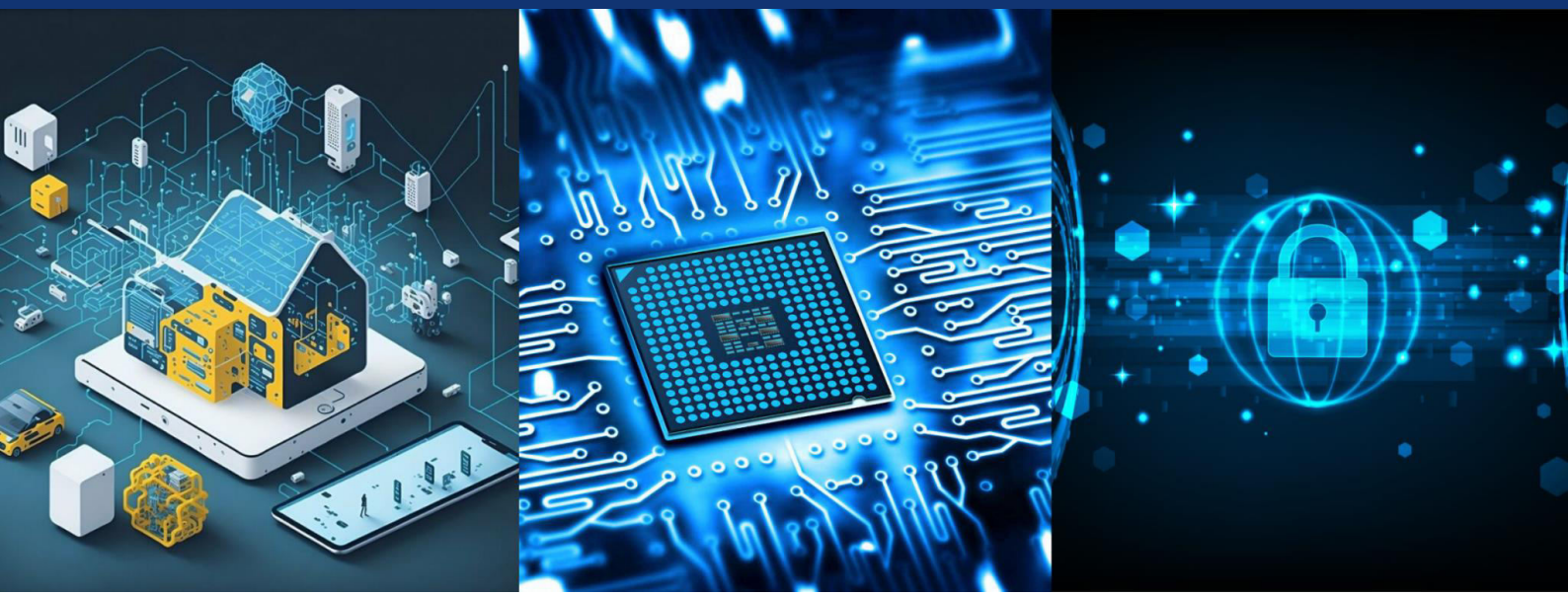
IJIRCCE

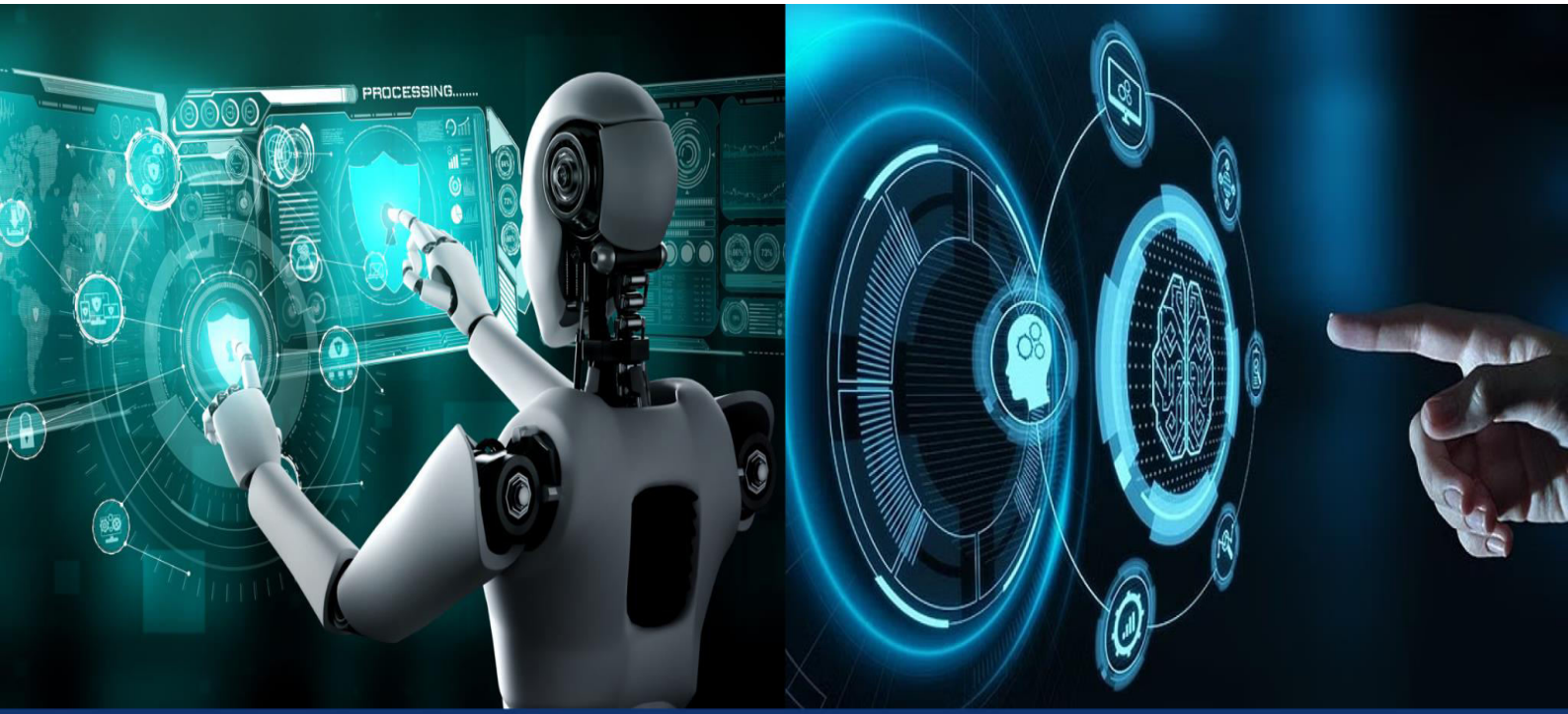
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 6, June 2026

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406014

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Customer Churn Analysis and Prediction

Dr. Kavyashree N, Ramya Shree D S

Associate Professor, Department of MCA, SSIT, Tumkur, Karnataka, India

4th Sem, Department of MCA, SSIT, Tumkur, Karnataka, India

ABSTRACT: Customer churn refers to customers discontinuing their relationship with a service provider. Predicting churn is essential for businesses because retaining existing customers is more cost-effective than acquiring new ones. This paper presents a Customer Churn Analysis and Prediction System developed using Machine Learning techniques. The system utilizes the IBM Telco Customer Churn dataset containing customer demographics, service subscriptions, and billing information. Various machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM are applied and compared. SMOTE is used to handle class imbalance, while SHAP explain ability helps interpret model predictions. The proposed system provides churn prediction, risk classification, automated reporting, and visualization dashboards. Experimental results show that XGBoost achieves the highest performance with a ROC-AUC score of 0.9367, enabling businesses to identify high-risk customers and implement proactive retention strategies.

KEYWORDS: Customer Churn, Machine Learning, XGBoost, SMOTE, SHAP, Predictive Analytics.

I. INTRODUCTION

In today's competitive business environment, customer retention has become one of the most important challenges faced by organizations across industries such as telecommunications, banking, insurance, e-commerce, and subscription-based services. Customers can easily switch between service providers due to the availability of multiple alternatives, competitive pricing, and improved digital services. This phenomenon, known as customer churn, directly impacts company revenue, profitability, and long-term business growth. Therefore, organizations focus on identifying customers who are likely to discontinue services and implementing effective retention strategies.

Traditional customer analysis methods mainly depend on manual reporting, historical observations, and basic statistical analysis. These approaches are often time-consuming and less effective when handling large-scale customer datasets. As organizations collect vast amounts of customer information daily, manual analysis becomes inefficient and may lead to delayed decision-making and inaccurate predictions. Hence, businesses require automated systems capable of analyzing customer behavior patterns and predicting churn probability efficiently. To overcome these limitations, this project introduces a "Customer Churn Analysis and Prediction System" using Machine Learning techniques. The primary objective of this system is to analyze customer information, identify churn-related patterns, and predict whether a customer is likely to leave the company. The proposed system helps organizations make data-driven decisions and improve customer retention strategies. The system processes customer datasets containing demographic details, service subscriptions, billing information, payment methods, tenure, and usage patterns. The collected data passes through several stages including data preprocessing, feature selection, exploratory data analysis, model training, prediction, and performance evaluation. Data preprocessing techniques such as handling missing values, encoding categorical variables, and feature scaling are applied to improve data quality and model efficiency. Various Machine Learning algorithms including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM are implemented and compared for churn prediction. These algorithms help classify customers into churn and non-churn categories based on their behavioral and service-related attributes. To improve prediction performance and handle class imbalance, SMOTE (Synthetic Minority Oversampling Technique) is applied during model training. The proposed system addresses several important business requirements such as: • Predicting customers who are likely to discontinue services.

- Identifying high-risk customers based on prediction probability.
- Analyzing factors influencing customer churn.
- Generating reports and visualization dashboards.
- Supporting customer retention and marketing strategies.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The need for such a system is increasing due to growing competition in the telecommunications and service industries. Organizations require efficient analytical systems capable of processing large amounts of customer information and generating accurate predictions. The proposed solution offers scalability and flexibility, making it suitable for deployment across various sectors including telecom companies, banking institutions, insurance organizations, online platforms, and retail businesses. The motivation behind developing the Customer Churn Analysis and Prediction System arises from challenges observed in traditional customer management processes. These challenges include high customer attrition rates, delayed identification of dissatisfied customers, revenue loss, and inefficient retention planning. By integrating Machine Learning techniques, the system aims to improve prediction accuracy, support faster decision-making, and enhance customer relationship management. In this paper, we present the design and implementation of the Customer Churn Analysis and Prediction System, including dataset collection, preprocessing techniques, feature engineering, model training, and performance evaluation. Different Machine Learning algorithms are analyzed and compared using performance metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Experimental results demonstrate that the proposed system achieves reliable prediction performance, helping organizations identify potential customer churn and implement proactive retention strategies..

II. LITERATURE REVIEW

Customer churn prediction has become one of the most widely researched areas in data mining and machine learning because customer retention plays a major role in improving organizational profitability and reducing revenue loss. Various researchers and organizations have proposed different analytical and predictive approaches to identify customers who are likely to discontinue services. Earlier studies mainly focused on statistical techniques, while recent research emphasizes Machine Learning models capable of handling large and complex customer datasets with improved accuracy.

Initial churn prediction systems were based on traditional statistical methods such as Logistic Regression and Linear Regression. These methods were preferred because of their simplicity, easy implementation, and interpretability. Researchers used customer demographics, billing information, and service usage patterns to predict churn probability. Although these approaches produced understandable results, they were limited in handling non-linear relationships and complex behavioral patterns present in real-world customer data. As datasets increased in size and complexity, traditional statistical models showed lower prediction performance compared to advanced machine learning techniques.

Decision Tree algorithms were later introduced for churn prediction because they provide rule-based classification and easier interpretation of customer behavior. Decision Trees divide customer records into multiple branches based on important features such as tenure, contract type, monthly charges, and payment methods. Researchers observed that Decision Tree models could effectively identify churn-related patterns and generate understandable classification rules. However, standalone Decision Trees often suffered from overfitting problems, leading to reduced generalization performance on unseen data.

To overcome the limitations of individual classifiers, ensemble learning techniques such as Random Forest and Gradient Boosting were introduced. Random Forest combines multiple Decision Trees and generates predictions using majority voting mechanisms. Studies showed that Random Forest algorithms improved prediction stability, reduced overfitting, and achieved better classification accuracy compared to traditional methods. Similarly, Gradient Boosting models sequentially build weak learners to minimize prediction errors and enhance overall model performance. These methods became highly popular in customer churn analysis due to their capability to handle high-dimensional customer datasets efficiently.

Recent studies have focused on advanced boosting algorithms such as XGBoost and LightGBM because of their high predictive accuracy and computational efficiency. XGBoost uses optimized gradient boosting techniques, regularization methods, and parallel processing to improve model performance and reduce training time. Many researchers reported that XGBoost achieves superior results in customer churn prediction compared to conventional machine learning models. LightGBM further improves efficiency by using histogram-based learning techniques and leaf-wise tree growth methods, making it suitable for large-scale datasets with faster execution speed.

Researchers also identified that customer churn datasets are often imbalanced because the number of non-churn customers is significantly higher than churn customers. This imbalance negatively affects prediction accuracy and may

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

bias models toward the majority class. To address this issue, oversampling and under sampling techniques were proposed. Among these techniques, SMOTE (Synthetic Minority Oversampling Technique) became one of the most commonly used approaches. SMOTE generates synthetic samples for minority class records and improves classification performance by balancing churn and non-churn datasets.

Feature engineering and exploratory data analysis have also played important roles in churn prediction research. Studies revealed that features such as customer tenure, monthly charges, contract type, internet services, payment methods, and total charges significantly influence customer churn behavior. Researchers used correlation analysis, visualization techniques, and feature importance methods to identify the most influential factors affecting churn prediction models.

Several comparative studies have evaluated the performance of multiple machine learning algorithms using metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Experimental results from previous research demonstrate that ensemble learning methods generally outperform traditional classification techniques in churn prediction tasks. Among all models, XGBoost and Random Forest consistently achieve higher predictive performance because of their ability to handle complex feature interactions and large customer datasets effectively.

Recent advancements in churn prediction research also emphasize automated reporting systems, visualization dashboards, and real-time prediction frameworks. Organizations increasingly require systems capable of generating business insights, customer risk categories, and retention recommendations automatically. Modern churn prediction platforms integrate data preprocessing, predictive modeling, visualization, and reporting into a unified analytical system that supports faster business decision-making.

Based on the analysis of previous research works, it is evident that Machine Learning techniques provide effective solutions for customer churn prediction. However, challenges such as dataset imbalance, feature selection, prediction interpretability, and model optimization still require continuous improvement. The proposed Customer Churn Analysis and Prediction System aims to address these challenges by integrating multiple machine learning algorithms, data preprocessing techniques, SMOTE balancing methods, and visualization modules into a single predictive framework.

III. METHODOLOGY

Our proposed Customer Churn Analysis and Prediction System consists of six stages: Data Collection, Data Preprocessing, Feature Engineering, Model Training, Churn Prediction, and Reporting & Visualization. The system analyzes customer behavior and predicts customers who are likely to leave the company.

A. Data Collection:

The system uses the IBM Telco Customer Churn dataset containing customer details such as gender, tenure, contract type, monthly charges, payment methods, and churn status.

Dataset Features:

Customer ID , Gender , Tenure , Contract Type , Monthly Charges , Payment Method , Churn Status

B. **Data Preprocessing**

Data preprocessing improves the quality of customer data before model training. Missing Value Handling: Missing values are identified and replaced using preprocessing techniques. Categorical Encoding: Categorical attributes are converted into numerical values using encoding techniques. Feature Scaling: Numerical features are normalized to improve model performance.

Class Balancing:

SMOTE is applied to balance churn and non-churn customer records.

C. Feature Engineering This stage identifies important customer patterns for prediction. Feature Selection: Important customer attributes are selected using correlation analysis.

Exploratory Data Analysis:

Graphs and charts are generated to analyze churn patterns and customer behavior.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

D. Model Training

Different Machine Learning algorithms are trained and compared.

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost Cross Validation: K-Fold Cross Validation is used to evaluate model performance.

E. Churn Prediction

The trained model predicts whether a customer is likely to churn.

Risk Classification:

- High Risk
- Medium Risk
- Low Risk

Performance Metrics:

- Accuracy
- Precision

F. Reporting and Visualization

The system generates dashboards and reports for analysis. Visualization: Charts and graphs display churn distribution and model performance.

Automated Reporting:

Prediction reports are generated to support customer retention strategies.

IV. RANDOM FOREST

Random Forest is one of the most widely used Machine Learning algorithms for classification and prediction tasks. It is an ensemble learning method that combines multiple Decision Trees to improve prediction accuracy and reduce overfitting problems. In the Customer Churn Analysis and Prediction System, Random Forest is used to classify customers into churn and non-churn categories based on customer behaviour and service usage patterns. The Random Forest algorithm helps the system to: • Predict customer churn accurately • Handle large customer datasets • Reduce overfitting problems • Improve overall prediction performance

Working Principle:

1. Multiple Decision Trees: The algorithm creates multiple Decision Trees using random subsets of customer data.
2. Feature Selection: Each tree selects random customer features such as tenure, monthly charges, and contract type for prediction.
3. Voting Mechanism: All Decision Trees generate predictions, and the final output is selected based on majority voting.
4. Final Prediction: The customer is classified as churn or non-churn based on the combined output of all trees.

V. XGBOOST

XGBoost (Extreme Gradient Boosting) is an advanced Machine Learning algorithm used for high-performance classification and prediction tasks. It improves prediction accuracy by combining multiple weak learners sequentially. In this project, XGBoost is used for accurate customer churn prediction because of its fast processing speed and high efficiency.

The integration of XGBoost enables the system to:

- Improve prediction accuracy
- Handle complex customer behavior patterns
- Reduce prediction errors
- Process large data

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406014

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. DATA VISUALIZATION

Data visualization is used to analyze customer behavior patterns and churn distribution using graphical representations. In the Customer Churn Analysis and Prediction System, charts and graphs are generated to identify important trends affecting customer churn.

Visualizations used in the system include:

- Churn Distribution Graph
- Correlation Heatmap
- Feature Importance Graph
- Confusion Matrix

These visualizations help organizations understand customer behavior and improve retention strategies..

V. ACCURACY IMPROVISATION

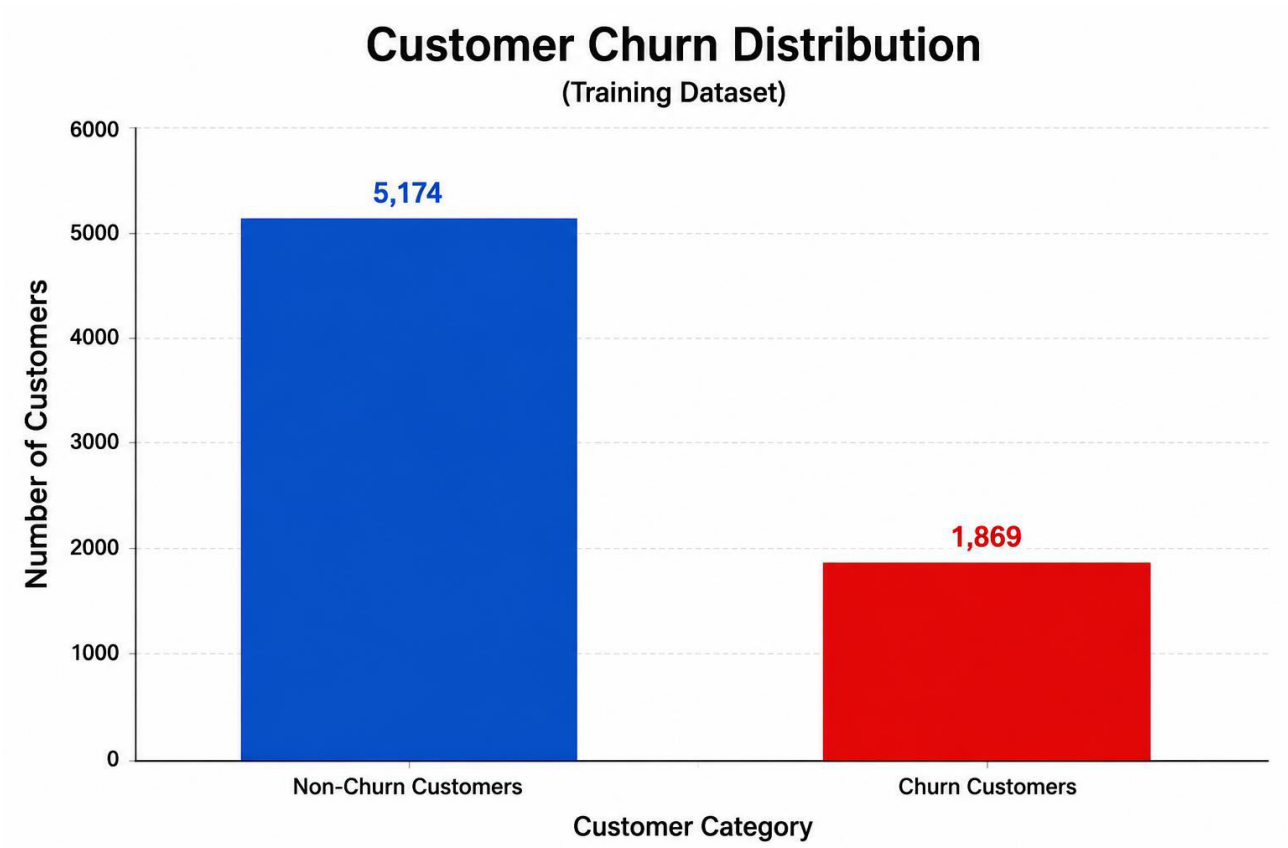


Figure 1: Dataset distribution showing number of samples used in training in each category.

This bar chart displays the number of training samples used in the Cypher Cam Master project. It includes four activity types: Normal Activity, Unattended Object, Fast Movement, and Suspicious Behavior.

Purpose:

- Ensures that the machine learning model is trained on balanced and diverse data.
- Helps prevent bias toward any one activity type.
- Improves classification accuracy during real-time surveillance.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406014

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Figure 2: Real-time churn prediction interface displaying customer risk alerts and prediction results.

This image shows the real-time customer churn prediction interface generated by the Customer Churn Analysis and Prediction System when high-risk customers are detected.

Purpose:

•Demonstrates real-time churn prediction and risk assessment functionality.
•Provides visual proof that the machine learning model successfully analyzes customer data and generates prediction alerts.
•Helps administrators identify high-risk customers instantly for retention strategies.
•Displays churn probability, customer risk level, and prediction results in an interactive dashboard.

VI. CONCLUSION

In this research, we proposed and developed a Customer Churn Analysis and Prediction System using Machine Learning techniques to identify customers who are likely to discontinue services. The system integrates multiple Machine Learning algorithms such as Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM for customer churn prediction and analysis.

The main objective of this project was to help organizations reduce customer attrition and improve customer retention strategies through predictive analysis. By applying data preprocessing, feature engineering, class balancing, and model optimization techniques, the system achieved reliable prediction performance and effectively classified customers into churn and non-churn categories.

The visualization and reporting modules further improve the effectiveness of the system by providing graphical insights, churn distribution analysis, and customer risk classification. This makes the proposed system suitable for

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

deployment in industries such as telecommunications, banking, insurance, e-commerce, and subscription-based services.

The project demonstrates that Machine Learning techniques can significantly improve customer churn prediction accuracy and support data-driven business decision-making. The proposed system also offers scalability and flexibility for future enhancements such as real-time prediction, cloud deployment, automated retention recommendation systems, and advanced customer analytics, making it a practical solution for modern customer relationship management.

REFERENCES

[1] Imani, Mehdi, et al. "Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning." *Machine Learning and Knowledge Extraction* 7.3 (2025): 105.
[2] Kuramannagari, Yuvakumar, et al. "A Comparative Analysis of Machine Learning Models for Customer Churn Prediction in Subscription-Based Businesses." *2025 2nd International Conference on Computational Intelligence and Computing Applications (ICCICA)*. IEEE, 2025.
[4] Rongala, Madan Kumar. "Comparative Evaluation of Machine Learning Models for Customer Churn Prediction in the Telecom Sector." *2025 International Conference on Computing Technologies (ICOCT)*. IEEE, 2025.
[5] Singh, Kiran Deep, et al. "Exploratory data analysis and customer churn prediction for the telecommunication industry." *2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*. IEEE, 2023.
[6] Dalvi, Preeti K., et al. "Analysis of customer churn prediction in telecom industry using decision trees and logistic regression." *2016 symposium on colossal data analysis and networking (CDAN)*. IEEE, 2016.
[7] Xiong, XuHai. "Online Learning Customer Churn Prediction Model Based on GA-XGBoost." *2024 10th International Conference on Computer and Communications (ICCC)*. IEEE, 2024.
[8] Vanitha, M. "Improved Bidirectional-Long Short-Term Memory for Customer Churn Prediction in the Telecom Industry." *2024 1st International Conference on Sustainability and Technological Advancements in Engineering Domain (SUSTAINED)*. IEEE, 2024.
[9] Murmanto, Imanuel Revelino, et al. "Application Design of Customer Churn Prediction Using Random Forest and XGBoost Algorithms for Telecommunication Industry in Indonesia." *2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*. IEEE, 2025.
[10] Shah, Jyoti Kunal, et al. "Unified AI Framework for Real-Time Customer Churn Prediction Using Behavioral and Event-Log Anomaly Signals." *2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC)*. IEEE, 2026.
[11] Zhou, H. "A Novel SHAP-Guided Bayesian Optimization Method for Telecom Customer Churn Prediction and Retention." *IEEE Trans. Network and Service Management* (2025).
[12] Bussey, Kathryn, Shatha Ghareeb, and Jamila Mustafina. "Bank Customer Churn Prediction: A Machine Learning Approach." *2025 18th International Conference on Development in eSystem Engineering*.
(DeSE). IEEE, 2025.
[13] Orhan, Busra Nur, and Engin Masazade. "Telecom Customer Churn Prediction and Root Cause Analysis Using Network Quality Metrics." *2025 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2025.
[14] Thanishka, B. Krishna, et al. "Machine Learning and Ensemble Prediction for Customer Churn Prediction Analysis." *2025 International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2025.
[15] Lu, Ning, et al. "A customer churn prediction model in telecom industry using boosting." *IEEE Transactions on Industrial Informatics* 10.2 (2012): 1659-1665.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details