International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05863
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Customer Churn Prediction using XGBoost

Sayali Tanpure..(Department of Data Science Dr. D. Y. Patil Arts, Commerce and Science College Pimpri)

Pooja Patil..(Department of Data Science Dr. D. Y. Patil Arts, Commerce and Science College Pimpri)

Abstract - Customer churn prediction is a critical task for businesses aiming to improve customer retention and maximize long-term profitability. This study proposes an advanced churn prediction model using XGBoost, an optimized gradient boosting technique designed for high performance on structured data.

The model is evaluated using multiple performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. Experimental results demonstrate that XGBoost significantly improves recall and AUC-ROC compared to traditional Gradient Boosted Decision Trees (GBDT), making it more effective in identifying potential churn customers.

The findings highlight the importance of using advanced ensemble methods for predictive analytics and provide actionable insights for developing targeted customer retention strategies.

Key Words: Predictive Analytics, Customer Retention, Ensemble Learning

1.INTRODUCTION

Customer churn refers to the phenomenon where customers discontinue their relationship with a company. Predicting churn in advance allows organizations to implement proactive retention strategies, reduce customer acquisition costs, and enhance customer lifetime value.

Traditional statistical methods often fail to capture complex, non-linear patterns in customer behaviour. Machine learning models, particularly ensemble methods, offer improved predictive capabilities.

Among these, XGBoost has emerged as a powerful algorithm due to its efficiency, scalability, and ability to handle structured datasets effectively.

This study focuses on leveraging XGBoost to improve churn prediction accuracy and enhance the identification of high-risk customers.

2.Data Collection and Preprocessing:

This study adopts a quantitative approach using supervised machine learning techniques to predict customer churn. The Telco Customer Churn dataset was utilized, which contains approximately 7000 customer records and 21 features, including demographic details, service usage patterns, and billing information. The dataset provides a comprehensive view of customer behaviour, making it suitable for predictive modelling.

Data preprocessing was performed to ensure data quality and consistency. Missing values were handled appropriately, and categorical variables were encoded into numerical form using label encoding techniques. The dataset was then divided into training and testing sets using a stratified split to maintain class distribution and ensure reliable model evaluation.

Model Architecture:

The churn prediction model was developed using XGBoost, an advanced ensemble learning algorithm known for its efficiency and ability to handle structured data. The model was trained using optimized hyperparameters, including learning rate, maximum depth, and number of estimators, to improve performance and generalization.

Model evaluation was conducted using multiple performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provide a comprehensive assessment of the model's predictive capability, particularly in handling class imbalance and identifying potential churn customers.

Model Evaluation

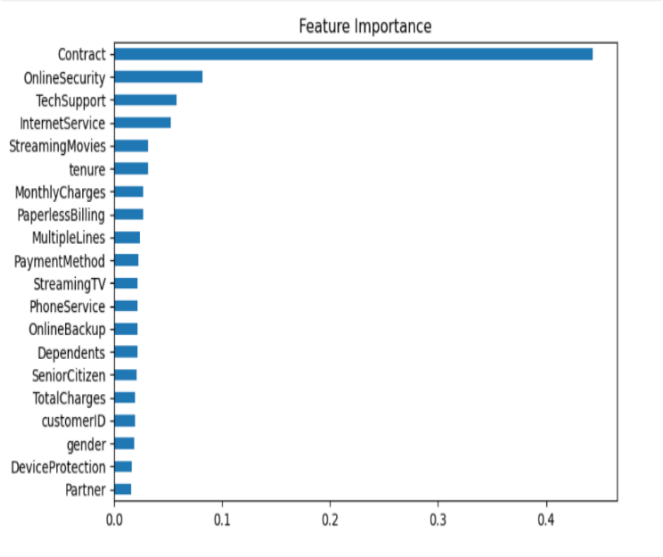
The performance of the churn prediction models was evaluated using key classification metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provide a comprehensive assessment of the model's ability to correctly identify churn customers while maintaining a balance between false positives and false negatives.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05863
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Metric	GBDT	XGBoost
Accuracy	0.7934	0.7530
Precision	0.6360	0.5239
Recall	0.5187	0.7593
F1-Score	0.5714	0.6200
ROC-AUC	0.8276	0.8390

Table 1: Model Performance Comparison

Chart



Result and Analysis:

The experimental results demonstrate that XGBoost outperforms GBDT in terms of recall, F1-score, and AUC-ROC, despite a slight decrease in accuracy and precision. The most notable improvement is observed in recall, which increases significantly from 0.5187 to 0.7593. This indicates that the XGBoost model is far more effective in identifying potential churn customers, which is crucial for minimizing revenue loss and enabling proactive retention strategies.

The higher AUC-ROC value further confirms that XGBoost has a stronger ability to distinguish between churn and non-churn customers across different thresholds. Although precision decreases slightly, this trade-off is acceptable in churn prediction, as identifying a larger number of at-risk customers is often more valuable than reducing false positives. Overall, the model demonstrates improved generalization and predictive performance.

The feature importance analysis (as shown in the figure) reveals that Contract is the most influential factor affecting customer churn, followed by features such as OnlineSecurity, TechSupport, and InternetService. Customers with shorter contract durations and limited service support are more likely to churn. Additionally, factors like tenure and monthly charges also contribute to churn behavior. These insights provide valuable guidance for businesses to design targeted retention strategies, such as offering long-term contracts and improving customer support services.

3.CONCLUSION:

This study demonstrates that XGBoost is a highly effective model for customer churn prediction, outperforming traditional GBDT in key performance metrics such as recall, F1-score, and AUC-ROC.

The significant improvement in recall highlights its ability to identify a larger proportion of at-risk customers, enabling businesses to implement timely and targeted retention strategies. While a slight reduction in precision is observed, the overall benefits of improved detection outweigh this limitation.

Future work can focus on further enhancing model performance through hyperparameter tuning, class imbalance handling techniques such as SMOTE, and exploring hybrid approaches combining machine learning with time-based models.

ACKNOWLEDGEMENT
We would like to express our gratitude to our college, faculty members, and project guide for their continuous support
and guidance throughout this research work

REFERENCES

1. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.
2. Kaggle (2018). Telco Customer Churn Dataset.
3. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques.
4. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python.
5. Friedman, J. H. (2001). Gradient Boosting Machine.