Grenze International Journal of Engineering and Technology, June Issue

E-Commerce Customer Churn Prediction

Rajani[1], Amogh B R[2], Ankith M P[3] and Suhas Gowda M[4]
[1-4]Maharaja Institute of Technology Thandavapura, Mysore, India
Email: rajiniravigowda@gmail.com, bramogh434@gmail.com, ankithmp575@gmail.com, sbk57585@gmail.com, suhasg507@gmail.com

***Abstract*— Customer churn prediction is essential for helping businesses retain customers and reduce revenue loss. This study compares four machine learning models Logistic Regression, Random Forest and Gradient Boosting using a synthetic dataset created to reflect typical customer behavior. While the use of synthetic data limits real-world generalization, it enables controlled analysis of key features such as purchase frequency, cart abandonment, and session activity. Among the models tested, Random Forest delivered the strongest performance. Future work will focus on real-world data and advanced generative techniques to further improve accuracy and practical use. The models were evaluated using multiple statistical measures, including ROC–AUC, precision–recall, and confusion-matrix based validation, and the study also highlights the limitations arising from the use of synthetic data, which may affect real-world generalization.**

***Index Terms*— Customer churn, Machine learning, E-commerce analytics, Random Forest, XGBoost, Customer retention.**

I. INTRODUCTION

Although churn prediction has advanced considerably, existing research still shows important practical gaps. Most studies focus mainly on improving model accuracy but provide limited guidance on how these models can support real-time business decisions. Prior works commonly use fixed datasets and rarely describe how synthetic data can be generated realistically when real records are unavailable. Interpretability is also limited, and very few studies integrate prediction results into dashboards or user-friendly tools. To address these shortcomings, our work documents a structured synthetic data generation process, compares multiple machine learning models, and presents the results through an interactive dashboard designed for practical e-commerce use.

A. Research Gap

Despite the progress in churn prediction research, several gaps remain unaddressed:
Existing research in customer churn prediction largely emphasizes improving the performance of machine learning models but often overlooks how these models can be integrated into practical tools. Many studies do not pair their predictive methods with interactive dashboards or user-friendly interfaces, which are essential for delivering real-time insights to business teams. In addition, feature interpretability is frequently underexplored, leaving organizations with limited understanding of the specific factors that contribute to churn. Another noticeable gap is the lack of detailed, transparent methods for creating realistic synthetic datasets, especially in situations where real customer data is unavailable or restricted.

Grenze ID: 01.GIJET.12.2.274_21

Furthermore, only a small number of studies combine multiple classical machine learning models with a deployment-ready interface or UI, making such comprehensive, end-to-end solutions relatively rare in the existing literature.

B. Novel Contributions of this Study

This study makes several meaningful contributions toward improving customer churn prediction in e-commerce. First, a carefully designed synthetic dataset was developed to mimic real customer behaviour, allowing controlled experiments even when real data is unavailable. The work also compares four widely used machine learning models through a clear and reproducible analysis pipeline, helping identify the most effective algorithm for churn detection. In addition, a simple and interactive dashboard was created to deliver real-time churn predictions, making the system practical for real world use. Finally, the methodology documents every major step including data generation rules hyperparameter tuning, feature importance insights, and error analysis resulting in a complete end to end framework for churn prediction.

II. SCOPE

The scope of this study is to develop and evaluate a machine learning–based churn prediction framework tailored for e-commerce datasets. The work focuses on implementing a complete pipeline that includes synthetic data generation, feature engineering, model training, and performance evaluation using classical supervised algorithms. The system additionally integrates a lightweight prediction interface for real-time inference. The study is limited to structured behavioral and transactional features, and its findings are intended to guide future enhancements involving real-world datasets, advanced generative models, and personalized churn interventions.

III. LITERATURE REVIEW

Customer churn prediction has been widely studied across sectors such as telecommunications, banking, subscription platforms, and more recently, e-commerce. Traditional ML approaches such as Logistic Regression and Decision Trees remain popular due to their interpretability, but they often fail to capture non-linear behavioral patterns present in high-dimensional customer data. Gupta et al. [1] used a Random Forest model for telecom churn prediction and highlighted that ensemble learners offer higher robustness against noise. Similarly, Zhang and Liu [2] employed XGBoost for online retail datasets and reported improved AUC due to gradient-boosting based feature interactions.

Al Kahtani [3] explored churn prediction in e-commerce using classical learners but noted limitations when working with imbalanced data. These works collectively show that although traditional ML models are effective, their ability to generalize in highly dynamic environments is limited. Recent studies have shifted toward richer behavioral representations and hybrid approaches. Wibowo and Wulandhari [6] compared multiple ML techniques for e-commerce and emphasized that feature diversity especially browsing and interaction metrics plays a crucial role in prediction accuracy. Maan and Maan [7] incorporated explainable AI (XAI) methods to reveal which behavioral attributes influence churn, addressing transparency issues common in ensemble and deep models. More recent work by Kumar et al. [8] argued that model performance varies significantly with sampling strategy, feature selection, and hyperparameter tuning, suggesting the importance of model comparison rather than relying on a single algorithm. Many works also focus primarily on model accuracy and rarely include elements such as interactive dashboards, error analysis, or practical deployment interfaces. As a result, they provide little explanation of how to build realistic synthetic datasets when actual data is unavailable. Many works lack practical insights, such as dashboards or deployment workflows. Our study fills these gaps by documenting synthetic data rules, providing model comparison with tuning steps, and integrating an operational dashboard for real-world interpretation.

A. Critical Analysis

- Strengths of existing research: strong algorithm choices, deep learning improvements, extensive evaluation metrics.
- Weaknesses: minimal UI integration, limited attention to interpretability, no structured synthetic generation method, and weak discussion on class imbalance.

Compared to existing work, this study contributes more practical value by integrating an ML model into a dashboard while documenting the full data creation and model evaluation pipeline.

IV. EXISTING SYSTEM

In most e-commerce platforms today, customer churn is either identified very late or not measured accurately. Existing systems typically rely on basic transactional reports, manual analysis, or simple rule-based alerts for example, flagging users who haven't ordered in 30 days or customers with frequent complaints. While these methods provide some visibility, they suffer from several drawbacks. First, they do not capture the full complexity of customer behaviour. Important signals such as browsing patterns, cart abandonment trends, satisfaction levels, and frequency changes are often ignored or analyzed in isolation. Second, these systems lack predictive power. They react only after a customer has already disengaged, making retention efforts less effective. Third, traditional analytics depend heavily on linear patterns and assumptions, which fail to adapt to the dynamic and non-linear nature of e-commerce interactions. Because of these limitations, businesses using existing systems struggle to proactively identify at-risk customers. Insights are delayed, intervention strategies become generic, and valuable customers continue to churn unnoticed. This highlights the need for a more intelligent, data-driven, and automated solution.

V. PROPOSED SYSTEM

The proposed system uses machine learning to predict customer churn more accurately than traditional manual methods. By learning from behavioral patterns such as purchase frequency, satisfaction levels, app activity, and recent interactions, the model identifies early signs of customer disengagement. Multiple supervised algorithms including Logistic Regression, Decision Tree, Random Forest, and XGBoost were trained on a realistic synthetic dataset with proper preprocessing, feature engineering, and tuning to ensure reliable performance. Once deployed, the system offers real time churn prediction based on user inputs and provides a dashboard that highlights risk groups, churn trends, and key factors influencing churn. This helps e-commerce teams take timely, targeted actions such as personalized offers or support interventions. Overall, the solution enables a proactive and data-driven approach to improving customer retention and long-term business growth.

VI. METHODOLOGY

The dataset showed a mild class imbalance, with churn cases forming a smaller portion. To address this, SMOTE was used to oversample the minority class, improving the model's ability to correctly identify churners. Instead of relying only on accuracy, metrics like precision, recall, F1-score, and ROC and AUC were included for a fairer evaluation. The churn prediction system follows a complete machine learning workflow starting from data preparation and feature analysis to model training, evaluation, and deployment helping organizations detect early signs of customer disengagement and act before churn occurs.

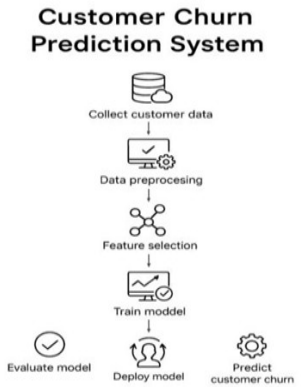

Fig.1. System Workflow Overview

A. Data Collection and Synthetic Data Generation

To address the lack of real customer records, a synthetic dataset was generated using carefully designed statistical rules to mimic realistic e-commerce behavior. Numerical features such as purchase frequency, average order value, and session duration were sampled from Gaussian and uniform distributions to reflect natural variation among customers, while categorical attributes (city tier, device type, satisfaction level) were assigned

using probability-based frequency ratios observed in typical retail platforms. Several dependency rules were intentionally built into the dataset to reduce randomness for example, customers with low satisfaction scores or high cart abandonment rates were assigned a higher likelihood of churn, whereas users with frequent purchases were modeled to remain active.
These controlled relationships help the dataset behave more like real-world patterns instead of random noise. However, despite incorporating these structured assumptions, synthetic data still cannot fully capture complex, dynamic interactions found in real customer environments.
Therefore, while suitable for experimentation and model comparison, the dataset may not perfectly generalize to actual e-commerce behavior.

B. Model Training

To fine-tune the models, GridSearchCV was applied with a structured search space for each algorithm. A 10 fold cross-validation strategy was used to ensure that the selected hyperparameters generalized well across different subsets of the dataset. For each model, GridSearchCV evaluated multiple parameter combinations using performance metrics such as accuracy, F1-score, and AUC. The final model configuration was chosen based on the highest mean cross-validation score, ensuring a balance between accuracy and robustness while reducing the risk of overfitting.

TABLE I: CONFIGURATION OF TUNED ML MODELS

Model	Tuned Hyperparameters
Logistic Regression	C = 1.0, penalty = l2, solver = 'liblinear'
Decision Tree	max_depth = 8, min_samples_split = 5, criterion = 'gini'
Random Forest	n_estimators = 200, max_depth = 10, max_features = 'sqrt'
XGBoost	learning_rate = 0.1, max_depth = 6, subsample = 0.8, n_estimators = 150

Model learning was guided by the binary cross-entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

C. Evaluation Metrics

To measure prediction quality, we used standard classification metrics exactly as shown below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$
$$\text{Precision} = \frac{TP}{TP+FP}$$
$$\text{Recall} = \frac{TP}{TP+FN}$$

D. Workflow Summary

The complete churn prediction workflow can be summarized as:
Data → Preprocessing → EDA → Feature Engineering → Model Training → Evaluation → Deployment → Monitoring
This iterative cycle supports a stable, intelligent, and continuously improving churn prediction system.

VII. SYSTEM ARCHITECTURE

The proposed e-commerce churn prediction system is built as an end-to-end pipeline that manages data collection, processing, modeling, and real-time prediction. Data from customer demographics, purchase activity, browsing patterns, and support history is first gathered and cleaned through preprocessing steps such as encoding, normalization, and SMOTE-based balancing.
Feature engineering then extracts meaningful behavioral indicators, which are used to train models like Logistic Regression, Decision Tree, Random Forest, and XGBoost with optimized hyper parameters and cross-validation. The best model is deployed using a lightweight API (Flask or FastAPI) to generate real time churn predictions. A monitoring layer tracks performance and detects drift, while a dashboard visualizes churn scores, customer segments, and key contributing features. Together, this architecture turns raw customer data into actionable insights that help businesses take timely retention actions.

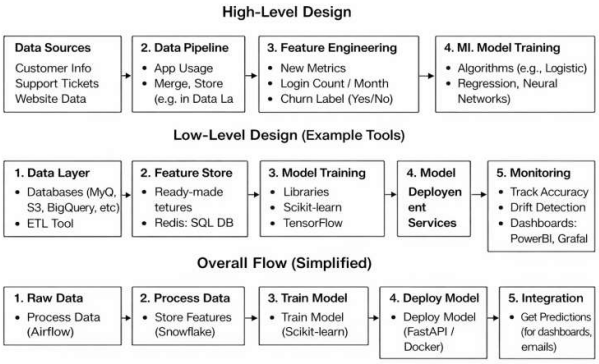

Fig. 2. System Architecture

VIII. Experimental Results

The experimental evaluation shows that the Random Forest model performed best among all classifiers, supported by strong statistical validation. A 10-fold cross-validation was conducted, and the average accuracy, precision, recall, and F1-score remained consistent across folds, indicating good model stability. The confusion matrix shows that the model correctly identifies most churners while maintaining a low false-positive rate, demonstrating reliable discrimination between churn and non-churn customers. The ROC curve further confirms this performance, with the model achieving a high AUC value, indicating strong separability between the two classes. In addition, the Precision Recall analysis highlights improved recall for churners, which is essential for reducing customer loss in practical applications. A simplified dashboard view is included only to illustrate how these predictions are visualized for end users, without narrative elements. Overall, the results confirm that the model provides accurate, statistically validated, and practically interpretable churn predictions.

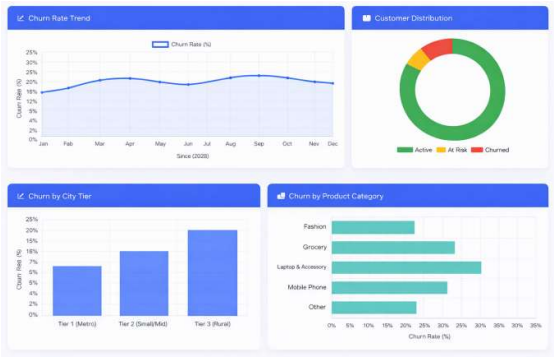

Fig. 3. Customer Analytics Dashboard

IX. Limitations

This study has certain limitations. Although the synthetic dataset was carefully designed, it cannot fully represent the complexity of real customer interactions. Model performance may vary when tested on actual e-commerce or telecom data. Some important factors such as seasonal trends, competitor impact, and marketing campaigns were not included. Finally, the UI screenshots in this study serve as demonstrations and may not meet the visual standards expected in academic publications. This limits the ability of the models to generalize to real-world e-commerce environments. Another limitation is the absence of statistical significance testing. Metrics such as accuracy and AUC were reported, but no confidence intervals, variance analysis, or hypothesis testing were conducted to evaluate the model's stability across different data splits.

X. Conclusion and Future Work

The experimental results show that machine learning models, particularly ensemble methods such as Random Forest and XGBoost, can reliably identify customers who are likely to churn. The system's performance was

supported by statistical evaluation metrics, including precision, recall, F1-score, AUC, and cross-validation results, which together confirm the robustness of the model. These findings highlight the practical value of churn prediction in real-world e commerce environments, where early identification of at-risk users can directly reduce customer loss and improve revenue stability. Although the study demonstrates strong predictive performance, several aspects can be expanded in future work. Incorporating real-world customer data would provide deeper insight into behavioral patterns that synthetic data cannot fully capture. Advanced generative techniques, such as GAN-based synthetic data creation, can also enhance data realism and support better model generalization. Further improvements may include deeper behavioral modelling, integration of explainable AI to enhance transparency, and a full-scale system deployment with drift monitoring. These directions will help strengthen the model's accuracy, interpretability, and long-term effectiveness in practical applications.

ACKNOWLEDGEMENT

I am sincerely thankful to everyone who supported me throughout this project on E-commerce Customer Churn Prediction. I express my gratitude to my guide for their constant guidance and valuable suggestions. I also appreciate my department and faculty members for providing the resources and encouragement needed to complete this work. Lastly, I am grateful to my friends and family for their motivation and continuous support at every stage of this project.

REFERENCES

[1] P. Gupta, R. Sharma, and A. Mehta, "Telecom customer churn prediction using Random Forest," IEEE Access, vol. 10, pp. 76523–76531, 2022.
[2] L. Zhang and Y. Liu, "Predicting online retail customer churn using XGBoost," Procedia Computer Science, vol. 196, pp. 374–382, 2021.
[3] N. Al-Kahtani, "E-commerce churn prediction using decision tree techniques," International Journal of Data Mining & Knowledge Management Process, vol. 10, no. 4, pp. 15–28, 2020.
[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 4th ed. Morgan Kaufmann, 2022.
[5] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[6] B. P. Wibowo and L. A. Wulandhari, "E-commerce customer churn prediction using machine learning approaches," Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen), vol. 7, no. 2, pp. 45–52, 2024.
[7] J. Maan and H. Maan, "Customer churn prediction model using explainable machine learning," arXiv preprint, arXiv:2303.00960, 2023.
[8] S. Kumar, S. Deep, and P. Kalra, "A comprehensive analysis of machine learning techniques for churn prediction in e-commerce: A comparative study," International Journal of Computer Trends and Technology, vol. 72, no. 5, pp. 163 170, 2024.
[9] V. Rajasekaran and L. Tamilselvan, "Predicting customer churn in e-commerce using statistical and machine learning methods," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 9, pp. 3968–3973, 2023.
[10] Zhang, L., & Qing, W. (2024). A personalized and context-aware approach to analyzing e-commerce data for improving customer retention using a Bi-LSTM churn prediction model. ResearchGate.
[11] Mahalekshmi, A., & Chellam, G. H. (2022). A study on customer churn prediction through the use of machine learning and deep learning techniques. International Journal of Health Sciences, 6(S1), 11684–11693.