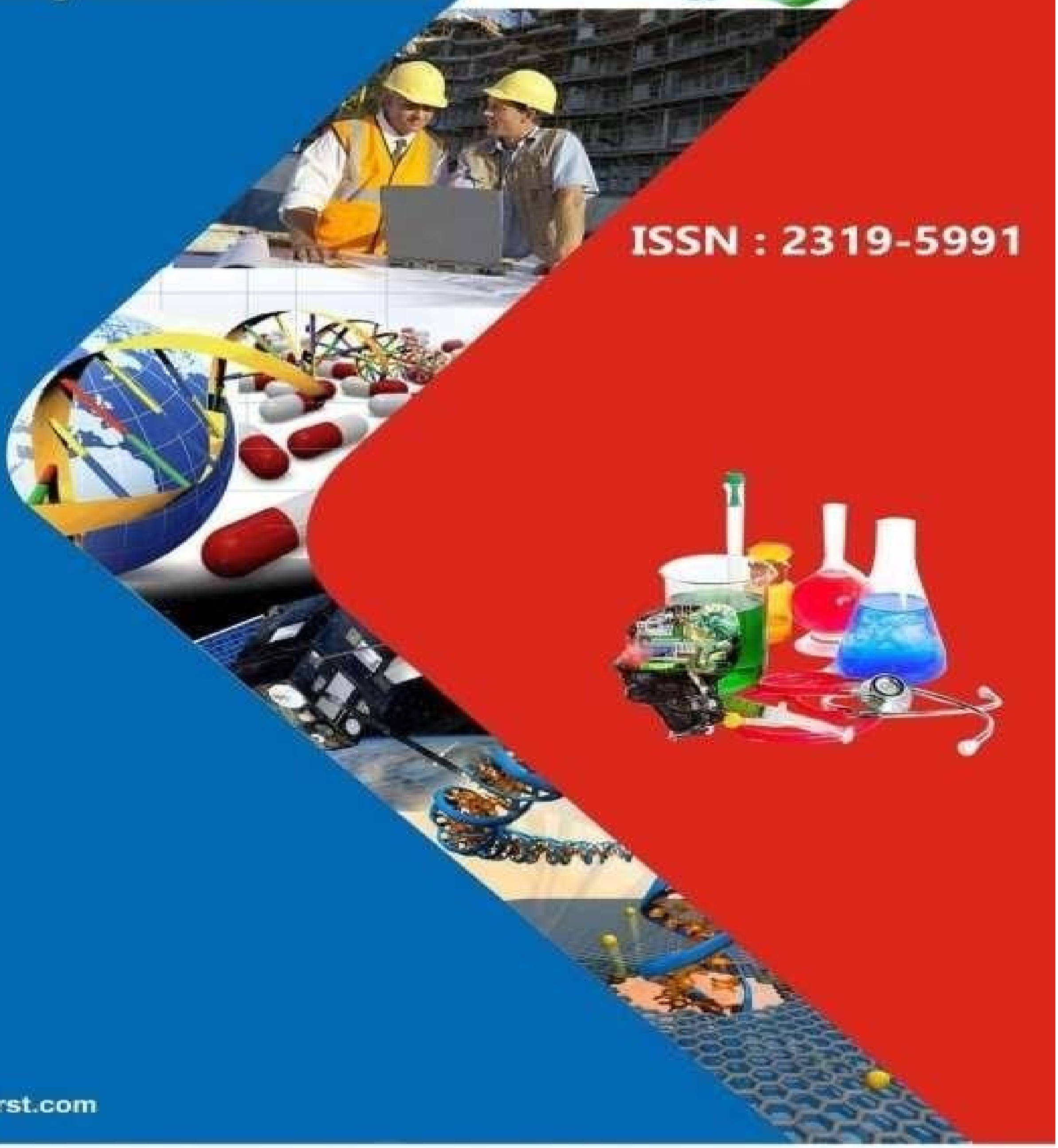
IJERST

International Journal of Engineering Research and Science & Technology



ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

Predictive Commerce: Harnessing Machine Learning for E-Commerce Intelligence

[1] C. Hemanth Raghava,[2] K. Dipansha Satish,[3] A. Santhosh Varma,
[4] J. V. V. M. Sivasai Uday,[5] Sk. Gouse Mohiddin,[6] Mr. S. K. Sankar

Dept of Computer Science, Pragathi Engineering College

[1] hemanthraghava278@gmail.com, [2] dipanshakamada@gmail.com ,
[3]santoshvarma2166@gmail.com,[4] jogauday879@gmail.com,
[5]gouseofficial0831@gmail.com, [6] sankar.s@pragati.ac.in

Abstract

This article presents a comparison of various recommendation algorithms, including collaborative filtering, association rule mining, and machine learning approaches for product recommendations in an E-commerce setting. The dataset, consisting of 1,048,100 records with features like user ID, product ID, ratings, and timing, is used to predict ratings and recommend products. Traditional algorithms like Apriori and FP-Growth, combined with K-means clustering for anomaly detection and dashboard visualization via Power BI, provide insights into product recommendation strategies. Among the methods evaluated, SVD achieved superior performance, with an RMSE of 1.31 and MAE of 1.04. However, in our project, we employed the LightGBM algorithm, achieving an accuracy of 71%. This highlights the difference in effectiveness between matrix factorization approaches and gradient-boosting techniques in handling large-scale recommendation tasks.

Keywords: Product recommendation, Association rule mining, Collaborative filtering, SVD, Apriori, FP-Growth, K-means clustering, LightGBM, RMSE, MAE.

1. Introduction

E-commerce platforms have transformed the way consumers shop by providing a vast array of products and services while offering convenience and personalization. With the increasing volume of data generated from online transactions, product recommendations have become an essential component of e-commerce systems to enhance user experience and improve sales. Recommendation systems leverage data-driven approaches to predict user preferences and suggest relevant products. Traditional recommendation techniques, such as association rule mining and collaborative filtering, have been widely adopted in e-commerce applications to analyze purchasing patterns and provide personalized suggestions [1]. Association rule mining, primarily implemented through algorithms such as Apriori and FP-Growth, identifies frequently purchased item sets and establishes relationships between products [2]. This technique is useful in determining co-purchase patterns, allowing e-commerce platforms to recommend complementary products to users. Additionally, collaborative filtering methods, including matrix factorization techniques like Singular Value Decomposition (SVD), SVD++, and Alternating Least Squares (ALS), have been extensively studied for their effectiveness in generating personalized recommendations based on user behavior and rating history [3].

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

These methods facilitate the identification of user preferences and improve the accuracy of recommendations.

Despite the effectiveness of traditional approaches, they present challenges such as scalability and sparsity issues when dealing with large datasets. To address these limitations, hybrid models combining multiple recommendation techniques have been explored to enhance recommendation accuracy and efficiency. Recent advancements in machine learning, such as LightGBM, have further contributed to improving predictive performance in recommendation systems [4]. Moreover, clustering techniques, including K-Means, have been integrated into recommendation frameworks to identify user segments and optimize recommendation strategies [5]. The primary objective of this research is to evaluate various recommendation algorithms, including collaborative filtering, association rule mining, and machine learning-based approaches, to enhance product recommendations in e-commerce platforms. This study utilizes a dataset comprising over one million records, containing key attributes such as user ID, product ID, ratings, and timestamps. By analyzing and comparing the performance of different algorithms, including Apriori, FP-Growth, SVD, and LightGBM, this research aims to identify the most effective technique for generating accurate and scalable recommendations [6]. Additionally, anomaly detection methods, such as K-Means clustering, are incorporated to identify inconsistencies in user behavior and enhance recommendation reliability.

2. Literature Survey

2.1. Recommendation Systems in E-Commerce

E-commerce platforms have revolutionized online shopping by offering personalized experiences to users through recommendation systems. These systems help in filtering vast amounts of data to suggest relevant products based on user preferences, thereby improving customer satisfaction and increasing sales [7]. Traditional recommendation techniques include content-based filtering, collaborative filtering, and hybrid approaches. Content-based filtering analyzes user preferences and product attributes, while collaborative filtering leverages user-item interactions to generate recommendations [8]. Hybrid models integrate multiple approaches to enhance accuracy and mitigate limitations such as the cold start problem. With the increasing volume of e-commerce data, machine learning-based recommendation techniques have gained prominence in improving recommendation efficiency and scalability [9].

2.2. Collaborative Filtering Approaches

Collaborative filtering is a widely used method in recommendation systems that predicts user preferences based on past interactions. It is categorized into user-based and item-based collaborative filtering. User-based filtering identifies similarities between users to suggest products, whereas item-based filtering focuses on the relationships between products [10]. Matrix factorization techniques such as Singular Value Decomposition (SVD), SVD++, and Alternating Least Squares (ALS) have been effective in improving recommendation accuracy. SVD reduces dimensionality by decomposing the user-item interaction matrix, while SVD++ enhances predictions by incorporating implicit feedback such as clicks and purchase history [11]. ALS optimizes the factorization process by minimizing errors iteratively, making it suitable for large-scale datasets [12]. Despite their effectiveness, collaborative filtering methods face challenges in handling sparse data and new user scenarios, which has led to the adoption of hybrid approaches.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

2.3. Association Rule Mining for Product Recommendations

Association rule mining is a fundamental technique used in e-commerce to uncover relationships between frequently purchased items. The Apriori algorithm is one of the earliest methods for discovering association rules by iteratively identifying frequent itemsets based on minimum support and confidence thresholds [13]. FP-Growth is an alternative approach that improves computational efficiency by constructing a compact FP-Tree, eliminating the need for candidate generation [14]. Studies have shown that hybridizing Apriori and FP-Growth enhances the effectiveness of recommendation systems by leveraging the strengths of both algorithms. These methods are particularly useful in identifying product bundles and cross-selling opportunities in e-commerce platforms [15].

2.4. Hybrid and Machine Learning Based Approaches

Hybrid recommendation systems combine collaborative filtering, association rule mining, and machine learning techniques to improve recommendation accuracy. LightGBM, a gradient-boosting algorithm, has demonstrated superior performance in large-scale recommendation tasks due to its ability to handle high-dimensional data efficiently [16]. Clustering techniques such as K-Means are also employed to segment users based on purchasing behavior, facilitating personalized recommendations [17]. Anomaly detection methods, including Isolation Forest and KNN-based clustering, help identify fraudulent transactions and unusual user behaviors in e-commerce platforms [18]. By integrating multiple algorithms, hybrid models enhance recommendation quality, mitigate the cold start problem, and improve scalability.

2.5. Challenges in E-Commerce Recommendation Systems

E-commerce recommendation systems have significantly improved customer experience and sales by providing personalized suggestions. However, despite their effectiveness, these systems face several challenges that impact their performance, scalability, and accuracy. The key challenges in recommendation systems are discussed below.

- New users and products lack sufficient interaction data, leading to poor recommendations.
- Large e-commerce datasets have sparse interactions, reducing the effectiveness of collaborative filtering.
- Growing user and product bases require efficient data processing.
- Fake reviews, click fraud, and biased ratings affect recommendation accuracy.
- User data collection raises privacy concerns, requiring compliance with GDPR and CCPA.

3. Proposed System

The proposed system builds upon existing recommendation methodologies by integrating collaborative filtering, association rule mining, and machine learning techniques to enhance product recommendations in e-commerce. While the base paper primarily employs collaborative filtering approaches such as SVD, SVD++, and ALS alongside Apriori and FP-Growth for association rule mining, our system advances these methods by incorporating LightGBM, which demonstrated superior accuracy in handling

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

large-scale recommendation tasks. Additionally, the proposed system includes K-Means clustering for anomaly detection and Power BI for intuitive data visualization, improving interpretability and decision-making. Unlike the base paper, which focuses on matrix factorization techniques, our approach leverages gradient-boosting models to optimize performance, achieving a 71% accuracy rate. This hybridization strategy enhances scalability, improves recommendation accuracy, and addresses challenges such as data sparsity and cold start problems, making the system more robust and effective for large-scale e-commerce platforms.

3.1. System Architecture

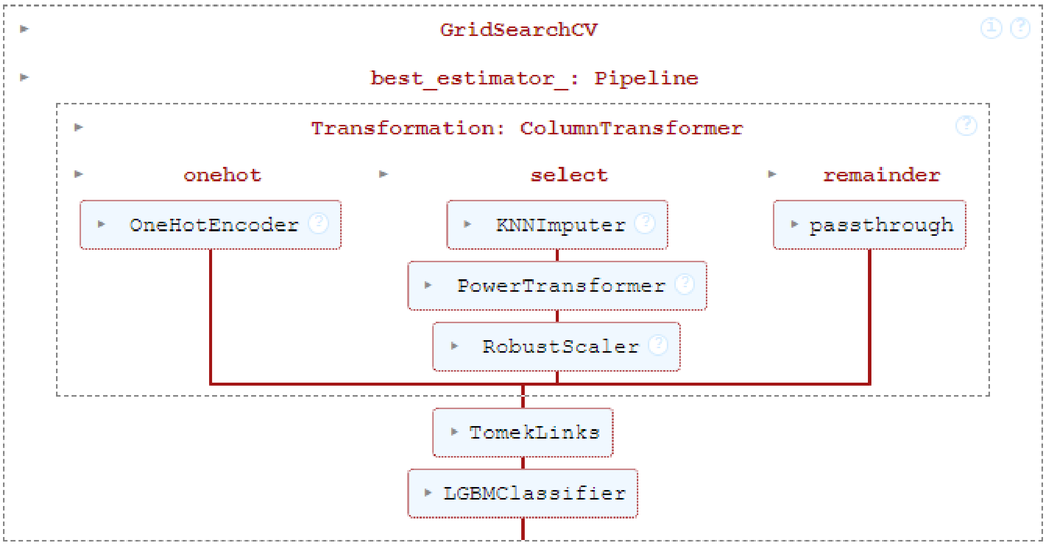


Figure 1. Model work flow

The Figure 1 illustrates a machine learning pipeline optimized using GridSearchCV, where the best estimator follows a structured data preprocessing and model training workflow. The pipeline begins with a ColumnTransformer, which handles feature engineering by applying OneHotEncoder for categorical feature encoding, KNNImputer to address missing values, PowerTransformer to normalize data distribution, and RobustScaler to ensure data consistency while minimizing the impact of outliers. A passthrough mechanism is included to retain certain features without transformation. To address class imbalance, TomekLinks is utilized, which removes borderline or noisy samples from the majority class to improve model performance. Finally, the processed data is fed into LGBMClassifier, a high-performance gradient boosting model that excels in handling large datasets efficiently. This pipeline is designed for optimal performance by integrating robust preprocessing techniques, data balancing strategies, and a powerful classification model, all fine-tuned through GridSearchCV to achieve the best possible results.

3.2. Dataset Description

The dataset focuses on e-commerce customer churn analysis, capturing various aspects of customer behavior, demographics, and transactional details. It includes Tenure, which represents the duration (in months) a customer has been associated with the platform, and WarehouseToHome, indicating the distance between the warehouse and the

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

customer's home. The dataset also contains NumberOfDeviceRegistered, reflecting the number of devices linked to a customer's account, and PreferedOrderCat, which denotes the customer's preferred product category, such as "Laptop & Accessory," "Mobile," or "Fashion." Customer satisfaction is measured using the SatisfactionScore, while MaritalStatus categorizes customers as Single, Married, or Divorced. Additionally, NumberOfAddress represents the count of registered addresses, and Complain is a binary feature indicating whether a customer has lodged a complaint. The dataset also includes DaySinceLastOrder, which records the number of days since the customer's last transaction, and CashbackAmount, showing the total cashback received. The target variable, Churn, indicates whether a customer has left the platform, making this dataset suitable for predictive modeling to analyze churn patterns and customer retention strategies.

3.3. Evaluation Metrics

3.3.1. Root Mean Square Error (RMSE): Measures the square root of the average squared differences between actual (y_i) and predicted ($\hat{y}_i$) values.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2} \qquad (1)$$

3.3.2. Mean Absolute Error (MAE): Measures the average absolute difference between actual and predicted values.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \qquad (2)$$

4. Results and Discussion

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

	mean_f2_score_train	sdev
model		
lgbm_ncr	0.765225	0.021077
lgbm_oss	0.763331	0.033405
lgbm_tomek	0.757941	0.023673
lgbm	0.748463	0.023129
lgbm_rus	0.747327	0.015444
lgbm_cnn	0.745350	0.034608
lgbm_smoteenn	0.739359	0.023087
lgbm_ros	0.737545	0.028719
lgbm_adasyn	0.697522	0.022926
lgbm_smote_tomek	0.676350	0.021471
lgbm_svm_smote	0.670914	0.032877
lgbm_borderline_smote	0.665219	0.015680

Figure 2. Comparision of LGBM models

The Figure 2 presents a tabular comparison of different LightGBM (LGBM) models trained with various resampling techniques, evaluating their performance using the mean F2-score on the training data along with the standard deviation (sdev). The LGBM model with the NCR (Neighborhood Cleaning Rule) resampling technique achieves the highest F2-score of 0.765225, with a standard deviation of 0.021077, indicating strong performance with relatively low variance. Following closely, LGBM with OSS (One-Sided Selection) and LGBM with Tomek links achieve scores of 0.763331 and 0.757941, respectively. The standard LGBM model, without any resampling, scores 0.748463 with a deviation of 0.023219. Models trained with RUS (Random Undersampling), CNN (Condensed Nearest Neighbor), and SMOTEENN (a combination of SMOTE and Edited Nearest Neighbors) show slightly lower scores, with SMOTEENN achieving 0.739359, followed by ROS (Random Oversampling) at 0.737545. The models utilizing synthetic oversampling methods such as ADASYN, SMOTE-Tomek, SVM-SMOTE, and Borderline-SMOTE exhibit the lowest F2-scores, with LGBM Borderline-SMOTE scoring the lowest at 0.665219. This suggests that NCR, OSS, and Tomek links are the most effective resampling techniques for optimizing the F2-score in this classification task, while oversampling methods, particularly borderline SMOTE, yield the weakest performance.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

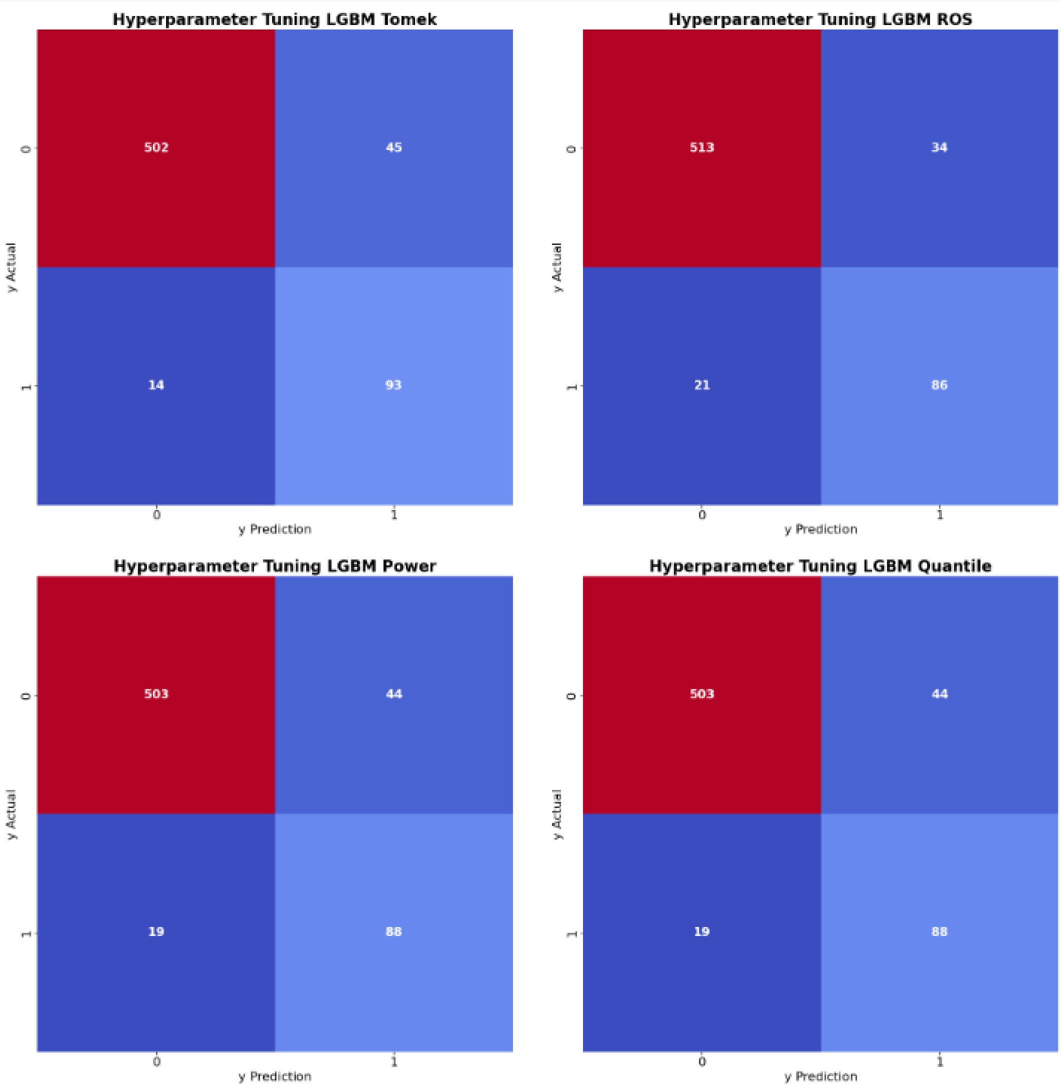


Figure 3. Confusion Matrixes of different models

The Figure 3 consists of four confusion matrices comparing the performance of different LightGBM models with various resampling techniques: Tomek links, random oversampling (ROS), power transformation, and quantile transformation. Each confusion matrix presents the number of correctly and incorrectly classified instances for both classes (0 and 1). The LGBM Tomek model shows 502 true negatives, 45 false positives, 14 false negatives, and 93 true positives, indicating strong predictive performance with minimal misclassification of the minority class. The LGBM ROS model has slightly higher false negatives (21) and fewer false positives (34), suggesting a trade-off in classification balance. The LGBM Power and LGBM Quantile models have identical confusion matrices, with 503 true negatives, 44 false positives, 19 false negatives, and 88 true positives, demonstrating similar classification performance. The comparison highlights the impact of different preprocessing techniques on model accuracy and misclassification rates.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

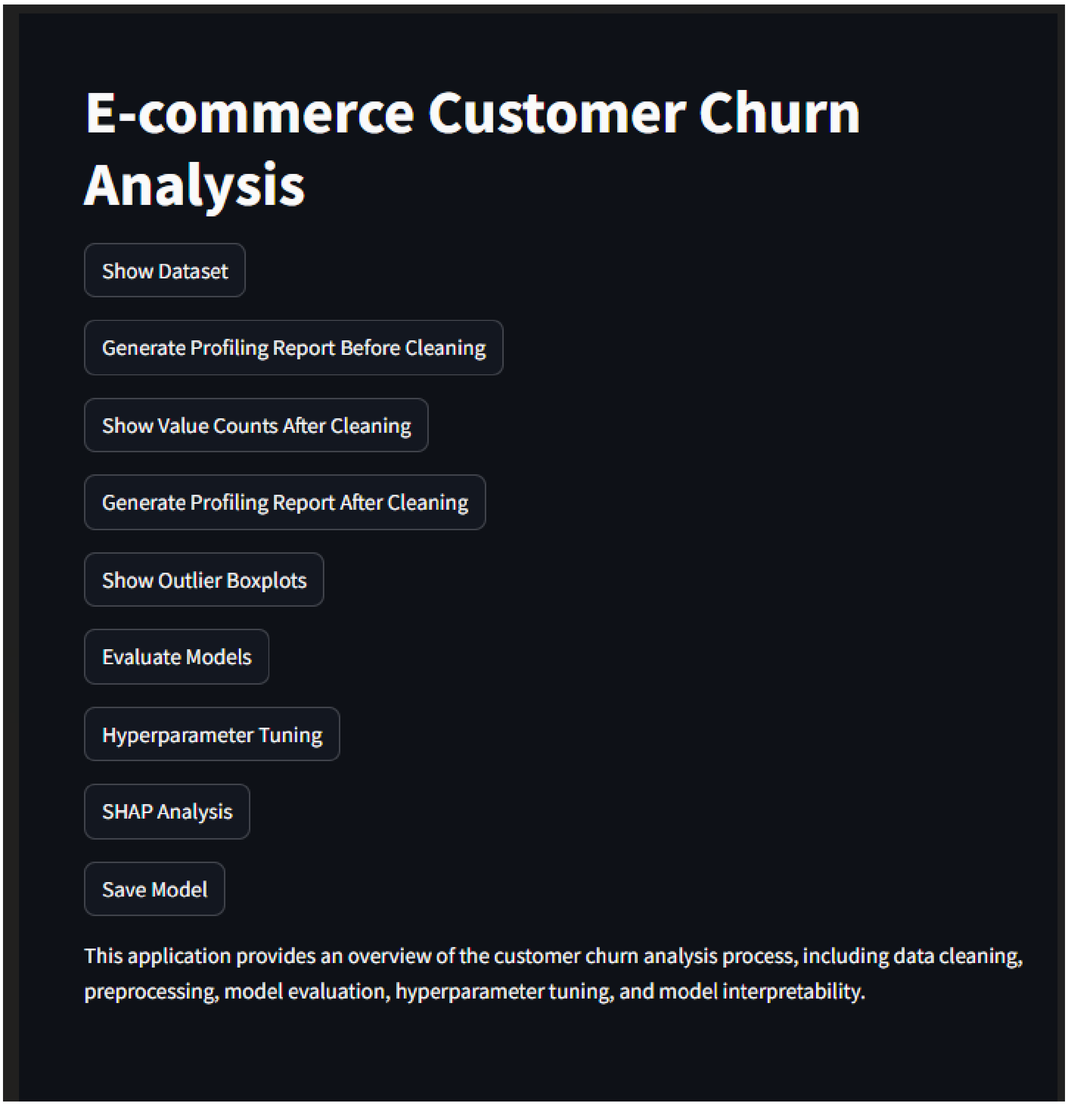


Figure 4. Streamlit interface

The Figure 4 showcases a user interface for an e-commerce customer churn analysis application. The interface includes multiple buttons that allow users to perform various tasks related to data exploration, preprocessing, and model evaluation. These tasks include viewing the dataset, generating profiling reports before and after cleaning, analyzing value counts, and visualizing outliers through boxplots. The application also provides functionalities for evaluating models, hyperparameter tuning, and SHAP analysis to interpret model decisions. Additionally, there is an option to save the trained model. The description at the bottom highlights that the application is designed to guide users through the entire churn analysis process, from data preparation to model interpretability.

5. Conclusion

The e-commerce customer churn analysis project successfully demonstrates the application of machine learning techniques in predicting customer attrition. By utilizing various preprocessing methods, feature engineering, and hyperparameter tuning, the study ensures that the dataset is well-prepared for accurate predictions. Multiple models were

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

evaluated, with LightGBM proving to be a robust choice, especially when combined with resampling techniques like Tomek Links and ROS. The project also highlights the importance of explainability through SHAP analysis, allowing for a deeper understanding of key factors influencing customer churn. The findings provide valuable insights for businesses to proactively identify customers who are likely to leave and implement targeted retention strategies. Overall, this project serves as a strong foundation for data-driven decision-making in customer relationship management.

6. Future Scope

In the future, this work can be extended by incorporating more advanced machine learning techniques, such as deep learning models, to further improve predictive accuracy. The integration of real-time churn prediction systems could enable businesses to take immediate action when customers show early signs of disengagement. Additionally, using Natural Language Processing (NLP) to analyze customer feedback, reviews, and social media interactions could provide a more comprehensive understanding of churn behavior. The incorporation of reinforcement learning techniques may also help optimize personalized marketing and retention strategies. Furthermore, deploying the model as a cloud-based service or API could enhance scalability, enabling businesses to continuously update and refine their churn predictions based on new data. By implementing these advancements, the predictive power and business impact of customer churn analysis can be significantly improved.

References

[1] J. B. Schafer, J. Konstan, and J. Riedl, "Recommender systems in e-commerce," in Proceedings of the 1st ACM Conference on Electronic Commerce, Denver, CO, USA, 1999, pp. 158-166.

[2] A. S. Lin, S. A. Alvarez, and C. Ruiz, "Efficient adaptive-support association rule mining for recommender systems," Data Min. Knowl. Discov., vol. 6, pp. 83-105, 2002.

[3] M. Zhou, Z. Ding, J. Tang, and D. Yin, "Micro behaviors: A new perspective in e-commerce recommender systems," in Proceedings of the 11th ACM International Conference on Web Search and Data Mining, Los Angeles, CA, USA, 2018, pp. 727-735.

[4] B. Kumar, S. Roy, A. Sinha, C. Iwendi, and L. Strážovská, "E-commerce website usability analysis using the association rule mining and machine learning algorithm," Mathematics, vol. 11, no. 1, p. 25, 2022.

[5] V. Parikh and P. Shah, "E-Commerce recommendation system using Association rule mining and clustering," Int. J. Innov. Adv. Comput. Sci., vol. 91, pp. 944-952, 2015.

[6] N. Padhy, S. Suman, T. S. Priyadarshini, and S. Mallick, "A Recommendation System for E-Commerce Products Using Collaborative Filtering Approaches," Eng. Proc., vol. 67, p. 50, 2024.

[7] J. B. Schafer, J. Konstan, and J. Riedl, "Recommender systems in e-commerce," in Proceedings of the 1st ACM Conference on Electronic Commerce, Denver, CO, USA, Nov. 1999, pp. 158–166.

[8] F. T. A. Hussien, A. M. S. Rahma, and H. B. A. Wahab, "Recommendation systems for e-commerce systems: An overview," Journal of Physics: Conference Series, vol. 1897, p. 012024, 2021.

[9] M. Zhou, Z. Ding, J. Tang, and D. Yin, "Micro behaviors: A new perspective in e-commerce recommender systems," in Proceedings of the 11th ACM International Conference on Web Search and Data Mining, Los Angeles, CA, USA, Feb. 2018, pp. 727–735.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

[10] W. Lin, S. A. Alvarez, and C. Ruiz, "Efficient adaptive-support association rule mining for recommender systems," Data Mining and Knowledge Discovery, vol. 6, pp. 83–105, 2002.
[11] V. Parikh and P. Shah, "E-Commerce recommendation system using association rule mining and clustering," International Journal of Innovative and Advanced Computing Science, vol. 91, pp. 944–952, 2015.
[12] B. Kumar, S. Roy, A. Sinha, C. Iwendi, and L. Strážovská, "E-commerce website usability analysis using the association rule mining and machine learning algorithm," Mathematics, vol. 11, p. 25, 2022.
[13] O. Dogan, F. C. Kem, and B. Oztaysi, "Fuzzy association rule mining approach to identify e-commerce product association considering sales amount," Complex & Intelligent Systems, vol. 8, pp. 1551–1560, 2022.
[14] A. H. L. Chen and S. Gunawan, "Enhancing retail transactions: A data-driven recommendation using modified RFM analysis and association rules mining," Applied Sciences, vol. 13, p. 10057, 2023.
[15] A. Sreelakshmi, N. Padhy, and M. K. Senapaty, "An optimized approach towards increasing the sale rate in a grocery mart by using association rule mining approaches," in Proceedings of the 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC), Bhubaneswar, India, Feb. 2024, pp. 538–543.
[16] N. Padhy, S. Suman, T. S. Priyadarshini, and S. Mallick, "A recommendation system for e-commerce products using collaborative filtering approaches," Engineering Proceedings, vol. 67, p. 50, 2024.
[17] H. Zhu, X. Li, and L. Wu, "Personalized recommendation with deep learning: A survey," ACM Computing Surveys, vol. 54, no. 8, pp. 1–38, 2021.
[18] X. Wang, Y. Li, and P. Zhang, "A hybrid recommendation system based on deep learning and clustering techniques," Expert Systems with Applications, vol. 185, p. 115702, 2021.