Optimizing Product Recommendations with Location-Based Data and Demographic Profiling Using CatBoost

Kirubahari.R[1],Sakthi Aashitha.N.S[2],Sakthi Jothi.T.S[3]

[1]*Associate Professor, Computer Science &Engineering, K.L.N. College of Engineering, Sivagangai,TamilNadu,India.*
[2,3] *Student, Computer Science &Engineering, K.L.N. College of Engineering ,Sivagangai,Tamil Nadu,India,*

Abstract - *Delivering highly personalized and contextually relevant product recommendations is essential for improving user satisfaction and business performance in today's competitive e-commerce environment. This work introduces an advanced product recommendation system that integrates content-based and collaborative filtering with location-based data, enhanced by the CatBoost gradient boosting algorithm. The system tailors recommendations based on individual user profiles, past interactions, and geographic context, ensuring alignment with user interests and preferences. By applying collaborative filtering, the system leverages the preferences of similar users to offer diverse and relevant product suggestions. Location-based data allows for recommendations that adapt to a user's geographic surroundings, offering location-specific products and services. CatBoost's gradient boosting capabilities improve recommendation accuracy by processing complex and categorical data, dynamically adjusting suggestions based on user feedback. This hybrid approach ensures that recommendations evolve with user preferences and contextual factors in real-time. The outcome is a personalized product recommendation pathway that significantly enhances user engagement and boosts business outcomes. This innovative system provides e-commerce platforms with a powerful tool for delivering highly targeted and effective product suggestions.*

Keywords: *Personalized Recommendations, Location Adaptation, Content-Based Filtering, Collaborative Filtering, CatBoost, Gradient Boosting.*

1.INTRODUCTION

In today's competitive e-commerce landscape, personalized and context-aware product recommendations are essential for improving user satisfaction and business performance. This system integrates content-based filtering, collaborative filtering, location-based data, and the CatBoost algorithm to deliver highly tailored product suggestions. Content-based filtering analyzes individual user profiles and past interactions, ensuring that recommendations align with specific preferences, while collaborative filtering leverages insights from similar users to diversify suggestions. Additionally, location-based data adjusts recommendations according to the user's geographic context, making them more relevant to their environment. The CatBoost algorithm enhances accuracy by efficiently handling complex and categorical data, ensuring that recommendations evolve in real-time to meet user preferences and market trends. Moreover, the system offers flexibility to businesses by allowing continuous customization of recommendation models based on specific product categories or seasonal trends. This comprehensive approach empowers e-commerce platforms to offer more relevant, personalized suggestions, thereby boosting user engagement and business success.

2. RETALED WORKS

In their 2021 study, Mustafa et al. [1] introduced "OntoCommerce," a hybrid recommender system aimed at overcoming significant challenges in e-commerce, particularly the cold-start and data sparsity issues. The system integrates ontology to systematically represent knowledge about customers and products, allowing for a better understanding of customer preferences and improving the calculation of user similarities. This is paired with sequential pattern mining (SPM), which analyzes user behavior over time to detect trends in interactions. By applying SPM to the outputs of collaborative filtering, OntoCommerce generates more personalized recommendations that are relevant to users even when historical data is limited. Experimental results demonstrated that this hybrid approach significantly outperformed traditional recommendation methods, providing accurate and tailored suggestions that enhance user experience in e-commerce platforms. Overall, OntoCommerce exemplifies how the combination of advanced techniques can facilitate better product discovery for customers, leading to increased satisfaction and engagement.

Jingyi Ding et al. [2] (2021) introduced a sales forecasting system in their paper titled "Sales Forecasting Based on CatBoost," which utilizes the CatBoost algorithm to enhance the accuracy of sales predictions in the retail sector. The authors trained their system using the

Walmart sales dataset, one of the most extensive datasets available for this purpose. By employing sophisticated feature engineering techniques, they aimed to improve both the accuracy of predictions and the efficiency of data processing.The results demonstrated that the CatBoost model significantly outperformed traditional forecasting methods such as Linear Regression and Support Vector Machine (SVM), achieving a remarkable Root Mean Squared Error (RMSE) of 0.605. A key advantage of the CatBoost-based model is its reduced requirement for fine-tuning compared to other algorithms, which enhances its adaptability across different custom datasets and broadens its practical applications in various retail contexts. This study underscores the effectiveness of the CatBoost algorithm in providing reliable and efficient sales forecasts, thus offering valuable insights for retailers looking to make data-driven decisions.

Park, Hong, and Cho[3] (2007) presented a location-based recommendation system in their paper "Location-Based Recommendation System Using Bayesian User's Preference Model in Mobile Devices," highlighting the increasing importance of location-based services (LBS) as mobile devices provide users with tailored information and services. As the volume of available information grows, users often struggle to identify options that align with their preferences and are timely accessible. The authors emphasize that personalized recommendation systems can significantly enhance user experiences through automatic filtering techniques. However, challenges arise on mobile devices due to limited screen size and resources, necessitating user interfaces that prioritize convenience and personalization. The paper discusses the "MovieLens Unplugged" project, which examines the usability of recommender systems on mobile devices with constrained displays and intermittent connectivity, underscoring the need for improved user interfaces. Additionally, Bayesian Networks (BN) are presented as effective tools for modeling user preferences by incorporating contextual information such as location, time, and user requests, allowing map-based personalized recommendation systems to better infer user needs and deliver relevant services. Ultimately, the study highlights the importance of integrating advanced algorithms and user-centric design in the development of location-based applications.

Kamble and Shaha [4] (2021) explored the enhancement of recommender system performance in their study titled "Product Recommendation System Using Machine Learning," focusing on the application of advanced data mining techniques, specifically Support Vector Machine (SVM) algorithms. They examined two common approaches: content-based recommending, which relies on item attributes, and collaborative filtering (CF), based on user interactions. However, the authors proposed an SVM-based model to achieve improved accuracy and adaptability in generating recommendations. Their experiments tested various product datasets, including groceries, beauty products, cold drinks, watches, and gourmet foods, to evaluate the SVM model's performance against traditional recommendation algorithms. The results revealed that the SVM-based system significantly outperformed conventional methods, demonstrating its effectiveness in capturing user preferences and delivering relevant recommendations. This study underscores the potential of integrating machine learning techniques into recommendation systems, ultimately enhancing user experience in e-commerce by providing a more personalized approach to product discovery.

R. Kirubahari and S. Miruna Joe Amali [5] (2024), in their study titled "An Improved Restricted Boltzmann Machine Using Bayesian Optimization for Recommender Systems," focus on addressing critical challenges that hinder the performance of recommender systems (RS), such as scalability, diversity, accuracy, and data sparsity. These issues are particularly pronounced in environments with rapidly increasing numbers of users and items, complicating neighborhood selection for recommendations. To mitigate these challenges, the authors introduce the Improved Restricted Boltzmann Machine (IRBM-BO), which enhances the traditional Restricted Boltzmann Machine (RBM) by incorporating Bayesian optimization to fine-tune hyperparameters like learning rate, momentum, and weight-cost. This integration allows the IRBM-BO to achieve better predictive accuracy through systematic adjustments of hyperparameters, as opposed to relying on time-consuming manual tuning. The authors validated their proposed method using well-known datasets such as MovieLens 100K, MovieLens 1M, and Netflix, assessing its performance through established metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The results of their experiments demonstrate significant improvements in both the effectiveness and accuracy of the recommender system, showcasing the potential of combining advanced machine learning techniques with Bayesian optimization to create more robust and efficient recommendation algorithms. By leveraging these innovations, the study emphasizes the potential for enhanced user experiences in systems where accurate and personalized recommendations are crucial for user satisfaction and engagement.

3. SYSTEM IMPLEMENTATION

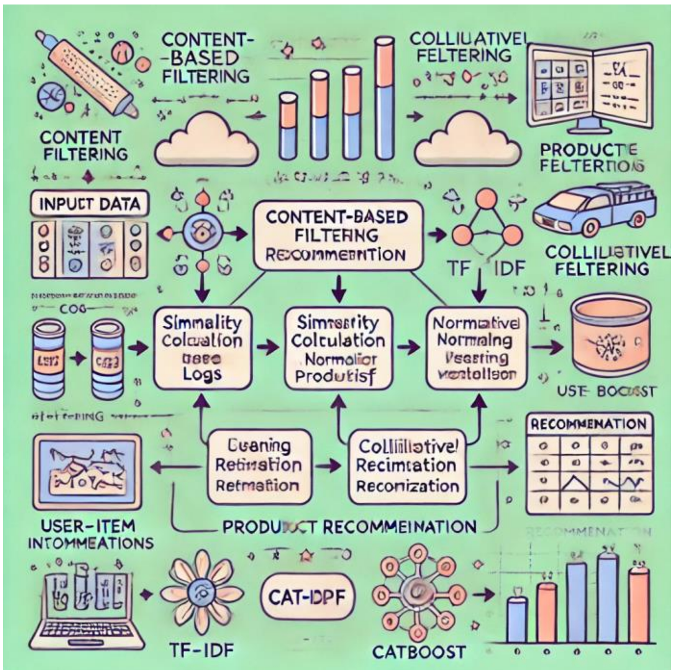

Figure 3.1 : ARCHITECTURE DIAGRAM

The diagrams illustrate the flow of our proposed Product Recommendation System, which integrates content-based filtering, collaborative filtering, and machine learning algorithms using CatBoost. The input data consists of user interaction logs, product descriptions, and relevant user information. This data undergoes a thorough preprocessing phase, including cleaning, normalization, and vectorization. The system employs TF-IDF for content-based filtering to analyze product descriptions, while collaborative filtering utilizes user-item interactions to enhance the recommendation process. The CatBoost algorithm then refines these recommendations by predicting product relevance based on historical user behavior and preferences.

3.1 Data Gathering

The input data for this recommendation system is sourced from an online retail dataset hosted on GeeksforGeeks. It includes key attributes like invoice numbers, stock codes, product descriptions, quantities purchased, customer IDs, and countries. After cleaning to remove missing values and unnecessary columns, only transactions with positive quantities are retained. This structured dataset forms the foundation for content-based and collaborative filtering models used in the system.

3.2 Data Preprocessing

Data preprocessing involves several steps aimed at optimizing the data for model training and recommendation accuracy. Initially, the dataset is filtered to remove any rows with missing descriptions or customer IDs. Subsequently, a TF-IDF Vectorizer is employed to convert product descriptions into numerical vectors, allowing for efficient similarity calculations. Additionally, the interaction matrix is created by pivoting the data to represent user purchases, enabling collaborative filtering techniques. This ensures that the data is well-prepared for the recommendation algorithms.

3.3 Similarity Calculation and Recommendation Generation

The recommendation process starts with calculating similarities between products using the content-based filtering method. For a given stock code, cosine similarities are computed based on the TF-IDF vectors, identifying products that are most similar in terms of description. For collaborative filtering, a user-item interaction matrix is utilized, which is reduced through Truncated Singular Value Decomposition (SVD) to extract latent features. This allows for personalized recommendations based on user purchase history and preferences. The system then combines the results from both methods to generate a comprehensive list of product suggestions.

3.4 CatBoost Model Integration for Recommendation Prioritization

The CatBoost algorithm is integrated into the system to enhance the recommendation process further. It takes the product recommendations generated from content-based and collaborative filtering methods and ranks them based on predicted probabilities of user interest. The model uses features such as product attributes and user interactions to provide an additional layer of personalization. This step ensures that the recommendations are not only relevant but also dynamically adjusted based on user behavior, making them more effective for engagement.

3.5 Product Recommendation and Visualization

The final stage involves presenting the top product recommendations to the user. The system ensures a minimum number of distinct recommendations by combining results from both filtering methods and supplementing with random selections if necessary. The recommendations are then displayed along with their corresponding probabilities, providing transparency about the suggested items. Additionally, a bar chart visualizes the top recommendations, enhancing user experience and facilitating better decision-making. This approach enables the system to provide tailored product suggestions that cater to individual user preferences and behavior effectively.

4. MODULES

4.1 Data Preprocessing Using TF-IDF Vectorization

The initial module focuses on preparing the dataset to optimize product recommendations. It starts by loading an online retail dataset that contains vital information, such as product descriptions, stock codes, customer IDs, and purchase quantities. The data undergoes thorough cleaning to eliminate any records with missing values in critical fields, which helps maintain an accurate representation of customer interactions. Relevant columns are retained, and a TF-IDF Vectorizer is applied to transform product descriptions into numerical formats. This normalization improves the calculation of product similarities and sets the groundwork for the subsequent recommendation algorithms. Furthermore, location data can be integrated during preprocessing to facilitate future filtering based on geographic relevance.

4.2 Content-Based Filtering using Cosine Similarity

This module implements a content-based filtering approach to generate tailored product recommendations. The system computes cosine similarities between product descriptions, allowing it to identify similar items based on their content. By incorporating location-based filtering, the system considers the user's geographical context when recommending products. This ensures that suggestions are not only aligned with the user's preferences but are also relevant to their specific region. When a user inputs a stock code, the system combines content similarities with location data to deliver personalized recommendations.

4.3 Collaborative Filtering Using Truncated Singular Value Decomposition(SVD)

In this module, collaborative filtering is utilized through an interaction matrix that captures user purchasing behavior. This matrix connects customer IDs with stock codes and their corresponding purchase quantities. It is processed using Truncated Singular Value Decomposition (SVD) to reveal hidden patterns in user preferences. Location-based filtering is also integrated, allowing the system to adapt recommendations based on the purchasing behaviors of users in similar geographic areas. This enhances personalization by considering both individual behavior and regional purchasing trends.

4.4 CatBoost Model Integration and Visualization of Recommendations

The final module focuses on utilizing the CatBoost classifier to enhance the recommendations derived from both content-based and collaborative filtering methods. Once initial recommendations are generated, the model ranks these suggestions according to predicted probabilities of user engagement, effectively managing categorical features. The outcomes are visualized through various techniques, such as bar charts, which display the top product recommendations along with their probability scores. This visualization assists users in quickly identifying relevant products, facilitating informed choices. The system is designed to adapt dynamically based on user input and feedback, ensuring ongoing improvements in recommendation quality.

5. RESULT

5.1 Product Names and Scores Table

The table presents product recommendations along with their relevance scores, indicating how well each product aligns with customer preferences. Higher scores, such as 0.8774 for the "PINK OWL SOFT TOY," suggest greater relevance, helping customers make informed choices. The table displays each product alongside its probability score showed in table 5.1.

Table 5.1 PRODUCT NAMES AND SCORES

PRODUCT NAME	SCORE
PINK OWL SOFT TOY	0.8774
BOX/12 CHICK & EGG IN BASKET	0.8735
SET OF 5 LUCKY CAT MAGNETS	0.8371
MULTICOLOUR 3D BALLS GARLAND	0.7913
CREAM SWEETHEART MINI CHEST	0.7833
LARGE ZINC GLASS CANDLEHOLDER	0.6233
PINK YELLOW PATCH CUSHION COVER	0.5972
REDRETROSPOT TRADITIONAL TEAPOT	0.2742
SWEETHEART BIRD HOUSE	0.2713
3 RAFFIA RIBBONS 50'S CHRISTMAS	0.2521

5.2 Graph of Product Recommendations

The horizontal bar chart visually represents product recommendations, with product names on the Y-axis and their corresponding scores on the X-axis, facilitating easy comparison. It features grid lines to enhance readability, clearly highlighting the top recommendations for better user insight. The probability scores are illustrated in the bar chart shown in **Figure 5.1.**

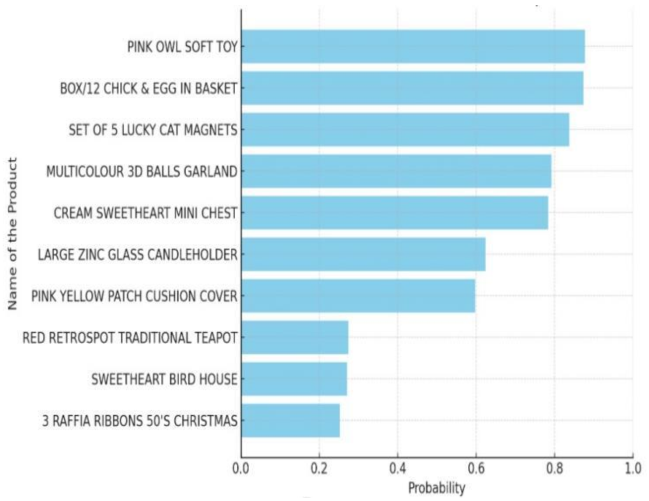

Figure5.1:TOP TEN PRODUCT RECOMMENDATIONS

5.3 Evaluation Metrics

The evaluation metrics indicate a reliable recommendation model with high accuracy (74.32%) and strong precision (72.05%), ensuring that most recommendations are relevant. Additionally, a good recall (68.91%) demonstrates the model's effectiveness in identifying a significant number of relevant products. The balanced F1 score (0.7046) and impressive ROC AUC (79.03%) further highlight its robust predictive capabilities.

Evaluation Metrics:

Metric	Value
Accuracy	0.7432
Precision	0.7205
Recall	0.6891
F1 Score	0.7046
ROC AUC	0.7903

Figure 5.2 : EVALUATION METRICS

6. CONCLUSION

A recommendation system for retail products has been developed that utilizes both content-based filtering and collaborative filtering techniques. These approaches enhance the shopping experience by suggesting products that align with customers' preferences based on their past behaviors and product characteristics. The data has been meticulously cleaned to remove incomplete or irrelevant information, allowing for a focused analysis of customers' buying habits. Content-based filtering employs TF-IDF (Term Frequency-Inverse Document Frequency) to analyze product descriptions, enabling the identification of similarities among products. This technique helps recommend items that closely match customers' interests. Meanwhile, collaborative filtering examines interactions among various customers to reveal patterns that can guide product suggestions, highlighting items that similar customers have appreciated.

The system allows users to input selected products and receive tailored recommendations. Results demonstrate the effectiveness of this system in providing relevant product suggestions, thereby enhancing customer satisfaction and potentially increasing retailer sales. Additionally, visual representations of the top recommendations are provided, showcasing the suggested products along with their likelihood of capturing user interest. As online shopping continues to expand, leveraging advanced machine learning techniques for recommendation systems proves to be an essential strategy for businesses aiming to improve customer experiences and foster loyalty

REFERENCES

[1].Ghulam mustafa, Naveed ahmad jhamat, Zeeshan arshad, Nadia yousaf(2024).”OntoCommerce: Incorporating Ontology and Sequential Pattern Mining for Personalized E-Commerce Recommendations” IEEE Explore , Issue on 18 March 2024 ,Vol.no:12.

[2].Jingyi Ding, Ziqing Chen, Li Xiaolong(2020).“Sales Forecasting Based on CatBoost”IEEE Explore , Issue on 2020, Vol.no:10.

[3].Moon-Hee Park, Jin-Hyuk Hong, and Sung-Bae Cho(2007).”Location-Based Recommendation System Using Bayesian User's Preference Model in Mobile Devices[3]”Springer, Issue 2007, Vol.no.4611.

[4].Nitin Kamble.Ajeenkya D. Y. Patil University(2021).”Product Recommendation System using Machine Learning”,SSRN, Issue on 11, November 2021, Vol.no:9.

[5].R. Kirubahari and S. Miruna Joe Amali(2024). "An Improved Restricted Boltzmann Machine Using Bayesian Optimization for Recommender Systems."Evolving Systems, Vol.15(3), pp.1099-1111 (https://doi.org/10.1007/s12530-023-09520-1).