An Intelligent Retail Analytics Platform Integrating Graph Convolution Networks and Explainable Machine Learning for Enterprise Monitoring and Predictive Insights

Borigam Ishwarya[1], T. Sanath Kumar[2*], Pedduri Rajesh[1], Vuppu Ruthvik[1], Singireddy Prashanth[1]

[1]UG Student, [2]Assistant Professor, [1,2]Department of Computer Science and Engineering (AI&ML)

[1,2]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India

*Correspondence: T. Sanath Kumar (sunny.554@gmail.com)

ABSTRACT

This study presents an intelligent retail analytics platform that integrates Graph Convolution Networks (GCNs) with explainable machine learning techniques to enhance enterprise-level decision-making. The system is designed to process heterogeneous retail data, including customer demographics, transactional records, and textual reviews, to generate actionable insights for business monitoring, evaluation, and strategic planning. By combining advanced predictive modeling with intuitive analytics, the platform enables a deeper understanding of customer behavior, product performance, and evolving sales trends across different market segments. The proposed architecture incorporates multiple machine learning models, including tree-based classifiers and regressors, along with a hybrid approach that integrates GCN with Natural Gradient Boosting (NGB). This combination significantly improves predictive performance in both classification tasks, such as customer rating prediction, and regression tasks, such as total sales forecasting. Furthermore, textual data is effectively processed using TF-IDF vectorization, enhancing the system's ability to capture customer sentiment, opinions, and feedback patterns. To ensure transparency and interpretability, the platform integrates explainable machine learning methods, allowing stakeholders to evaluate model outcomes using key performance metrics such as accuracy, precision, recall, F1-score, and error measures. Visual analytics tools, including confusion matrices, ROC curves, and correlation heatmaps, provide clear and interpretable insights for informed decision-making. Additionally, a web-based interface is developed to support secure user interaction, role-based access control, and real-time prediction capabilities. The platform also facilitates exploratory data analysis through interactive visualizations, enabling better understanding of sales patterns, demographic distributions, and market segmentation.

Keywords: Graph Convolution Networks (GCNs), Retail Analytics, Sales Forecasting, Customer Behavior Analysis, Natural Gradient Boosting, TF-IDF Vectorization.

1.INTRODUCTION

Retail analytics played a crucial role in modern corporate environments by enabling organizations to derive insights into customer preferences, product performance, and regional sales trends. With the rapid growth of e-commerce platforms and the expansion of diverse consumer segments, the retail landscape became increasingly dynamic and competitive.

Figure 1: The Benefits of Retail Data Analytics.

The emergence of omnichannel purchasing patterns further intensified the need for integrated and real-time data analysis. Organizations were required to continuously adapt to evolving customer expectations and the rise in digital interactions. As a result, the demand for intelligent, real-time analytical systems to support effective decision-making significantly increased. The increasing volume of transactional and customer-related data requires efficient processing of both structured and semi-structured information to support accurate decision-making. By leveraging advanced analytics, organizations can enhance demand forecasting, streamline inventory control, design targeted marketing strategies, and improve overall customer experience. Data-driven approaches allow retailers to identify hidden patterns, predict future trends, and respond proactively to market changes. Moreover, real-time analytics derived from well-prepared retail datasets empowers businesses to optimize pricing strategies, improve revenue planning, and deliver personalized services to customers. These capabilities provide a strong competitive advantage in a dynamic retail landscape. The major benefits of retail data analytics are clearly represented in Figure 1.

2. LITERATURE SURVEY

2.1 RELATED WORK

Marketfeed et al. [1] strategic growth: tea/coffee expansion, merger with Tata Global Beverages (Feb 2020), Starbucks JV, brand segmentation. Reports ROCE (~9.5%), nearly net debt-free, and implies TCP attractiveness from a fundamental lens. The retail industry has become highly informative through the collection of large amounts of transactional data every day [2]. These data are a gold mine when it comes to analyzing customer trends, managing stock, and increasing organizational effectiveness. To optimally utilize this wealth of data, health care organizations need not only strong analytical concepts and models, but analytical concepts and models that include exploratory, predictive, and prescriptive analytics [3]. Knowledge about how to use such data is important for being able to survive in the current business environment, more so for retail businesses.

The global retail market environment has become volatile in terms of demand and steadily increasing competition, which requires accurate demand forecasting and consumer segmentation as critical success factors [4]. The challenges that affect the retailers include the ability to forecast sales so as to avoid holding large stocks or running out of stock; identification of loyal customers who should be sustained as a way of reducing channel leakage; and finally, awareness of any new trends within the market so as to know which aspect to focus on while conducting marketing [5]. With changing business dynamics, where most decisions are made based on data, the use of ML, statistical modelling for data analysis, and advanced segmentation techniques have become a solution to these problems.

Exploratory data analysis or EDA has for a long time now been a staple of retail analytics where organizations can investigate large transactional datasets to find new patterns [6]. Information derived from EDA can include seasonality patterns for sales, transaction rates, and patterns of revenues [7]. Nevertheless, descriptive analytics form a basis, but this lacks the answer to creating necessary predictive reports and forecasts. This gap is filled by predictive modelling, which in turn helps the businesses to forecast or estimate the future performance and outcomes including the certain sales or the demand for certain products [8]. Among these methods, regression models are used most frequently for demand forecasting because of their ability to estimate such quantitative targets as constant growth rates, percentage increases, etc. However, the other critical aspect for the deployment of strategic marketing and engagement is customer behavior prediction [9].

RFM analysis, which stands for recency, frequency, and monetary, is an effective method for a customer's division according to his/her buying profile [10]. Studying customers as a valuable or loyal group, at-risk or potential to leave the company, or new, RFM analysis enables businesses to create subsequent marketing strategies for the above segments to ensure customer retention and satisfaction [11]. Combining predictive modelling and RFM analysis provides a synergy that effectively covers both probable selling and customer relations issues. The integration of regression models for demand forecasting with RFM-based customer segmentation presents a novel approach that simultaneously optimizes stock management and marketing strategies. By leveraging predictive analytics and behavioral clustering, this study provides a holistic view of retail dynamics. A major spear that grounds this study is that today's retailers require new knowledge from their inherent transactional data. Research has shown that many retailers face challenges in implementing exploratory data analysis (EDA), training predictive models, and segmenting customers effectively. These issues contribute to inefficiencies in inventory management and targeted marketing, highlighting the need for integrated analytical solutions [12].

This gap leads to cost waste in inventory, lack of proper segmentation in the market, and a general lack of responsiveness to market changes. This research is particularly driven by the possibilities that bring together more than one analytical approach to achieving comprehensive results. For instance, regression models can be used to predict sales trends, but partners cannot use this model as guide to solve the behavioral aspect of customer retention. Likewise, RFM analysis categorizes those vital customers into loyal and churner, but it gives no assistance in demand forecasting [13]. While previous studies have examined either demand forecasting or customer segmentation separately, few have explored their integration for comprehensive retail analytics [14]. This study fills this gap by combining predictive modelling with behavioural segmentation to provide a data-driven framework for strategic retail decision-making [14]. The information obtained from the results can contribute to the sound management decisions for inventory management, marketing, and customer relations activities as well as to increase profitability and reduce business costs.

3. PROPOSED SYSTEM

The proposed methodology defined a structured, data-driven framework for retail analytics to predict customer rating categories and total sales using both structured and unstructured data. The system followed a comprehensive pipeline that included data input acquisition, preprocessing, feature transformation, model training, and prediction stages. It integrated machine learning techniques with text processing to effectively analyze customer behavior and purchasing patterns.

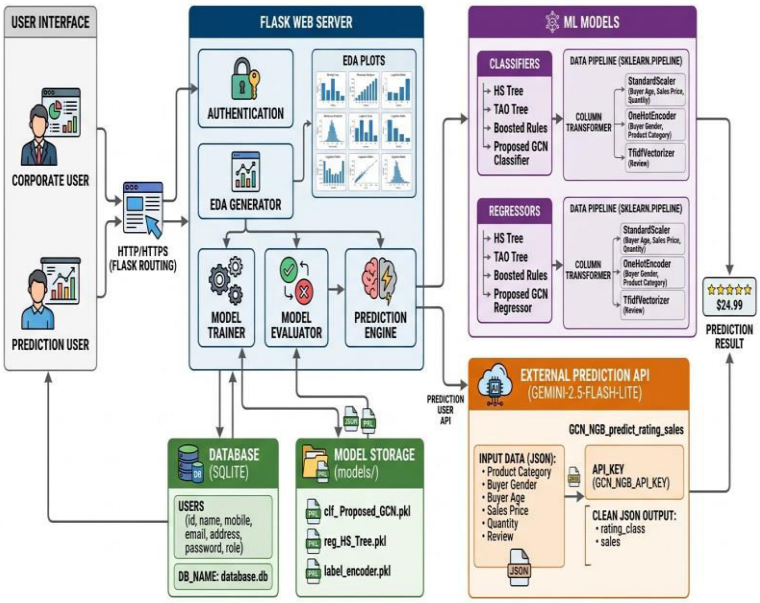

Figure 2: Proposed system architecture.

The framework was designed to handle heterogeneous data sources, including numerical, categorical, and textual inputs such as customer reviews. Additionally, comparative model evaluation and support for both real-time and batch predictions ensured robustness, consistency, and effective decision-making in retail environments as shown in figure 2.

User Interface (Web-Based Input System)

- Users provide inputs such as product category, buyer demographics, price, quantity, and textual reviews through a browser interface.
- The interface supports real-time prediction requests and administrative analytical operations via a responsive dashboard.
- All user interactions are transmitted securely to the backend server for processing and response generation.

Backend Processing Environment (Flask Server)

- Manages application routing, user authentication, and session handling to ensure secure data access.
- Coordinates the communication flow between the input modules, the machine learning pipeline, and the output rendering engine.
- Ensures seamless integration of data preprocessing, model execution, and result visualization.

Raw Input Data Representation

- Consists of structured attributes (age, price, quantity) alongside categorical and textual inputs.
- Includes customer reviews containing informal language, subjective opinions, and sentiment expressions.
- Serves as the primary multi-modal foundation for predictive modeling and trend analysis.

Data Preprocessing Module

- Performs normalization of numerical features using standard scaling techniques to ensure feature parity.
- Applies one-hot encoding for categorical variables to convert them into a machine-readable format.
- Utilizes TF–IDF vectorization to transform unstructured textual reviews into meaningful numerical representations.

Feature Transformation and Integration

- Combines numerical, categorical, and textual features using a robust column transformation pipeline.
- Ensures efficient handling of heterogeneous data types within a unified modeling framework to prevent data leakage.
- Produces a high-dimensional feature space optimized for deep learning and boosting algorithms.

Proposed Model: GCN with Natural Gradient Boosting

- Implements a hybrid approach combining Graph Convolution-based feature transformation with histogram-based gradient boosting.
- Enhances feature representation by capturing relational data structures while maintaining computational efficiency.
- Performs dual-output tasks: classification for Rating Prediction (Poor to Excellent) and regression for Sales Forecasting.

Baseline Machine Learning Models

- Includes HS Tree, TAO Tree, and BRC models to serve as benchmarks for comparative analysis.
- Each model independently processes the integrated features to generate predictions.
- Enables rigorous validation of the proposed hybrid model against traditional rule-based and tree-based approaches.

Prediction Output and Performance Analysis

- Generates discrete customer rating classes and continuous total sales values based on input attributes.
- Computes comprehensive metrics: Accuracy, F1-score (for classification) and MAE, RMSE, R2 (for regression).
- Visualizes results through confusion matrices, ROC curves, and scatter plots to support business intelligence and decision-making.

4. RESULTS ANALYSIS

This section presents the experimental results obtained from the implementation of the proposed retail analytics system. It explains the outcomes generated at different stages, including user interaction, data analysis, model evaluation, and prediction. The results demonstrate the effectiveness of ML models and the proposed GCN–NGB framework. Visual representations are used to clearly illustrate system performance and functionality.

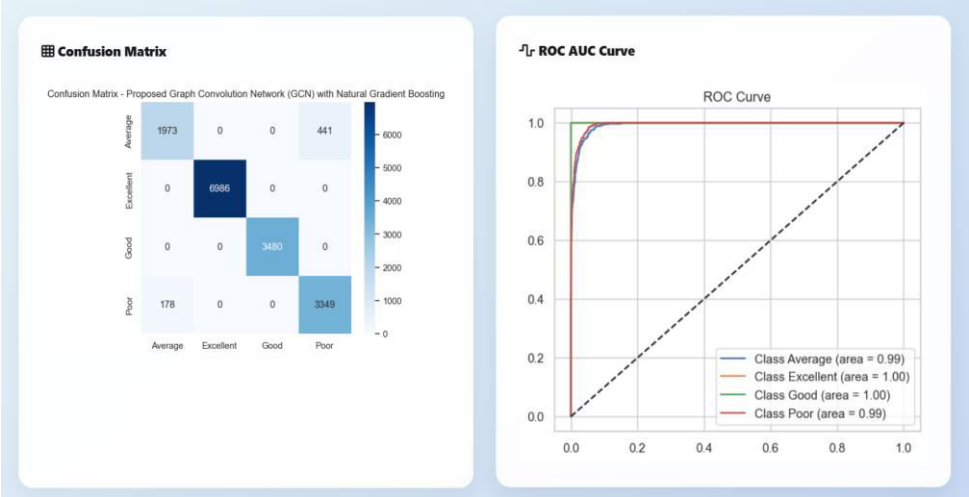

Figure 3 Confusion matrix and roc curve of rating target of various classifiers - GNC-NGB

Figure 3 GCN–NGB: The confusion matrix representing GCN–NGB shows the classification performance of the proposed model. The matrix exhibits a high concentration of correctly classified samples along the diagonal cells, indicating strong agreement between actual and predicted rating classes. The structure of GCN–NGB captures complex relationships among the sales features, resulting in highly accurate classification of rating categories within the corporate sales dataset.

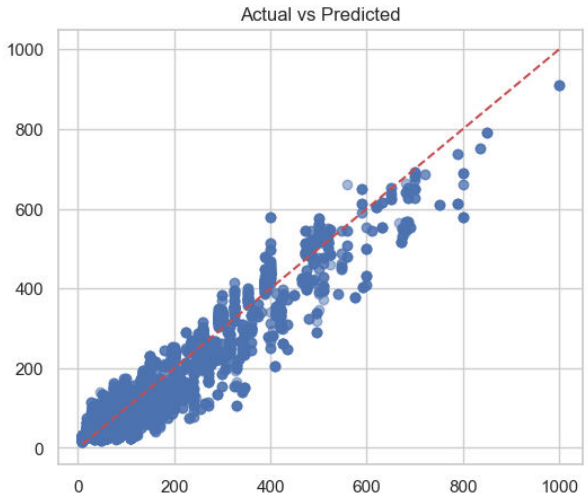

Figure 4. Scatter plots of total sales target of various regressors - GNC-NGB

Figure 4 GCN–NGB: The scatter plot representing GCN–NGB displays a dense concentration of prediction points along the diagonal reference line. This distribution indicates strong agreement between the actual and predicted rating values generated by the proposed model. The integration of graph-based feature representation with boosting optimization strengthens the model's ability to learn complex dependencies within the corporate sales dataset, resulting in highly consistent rating predictions compared to the other regressors.

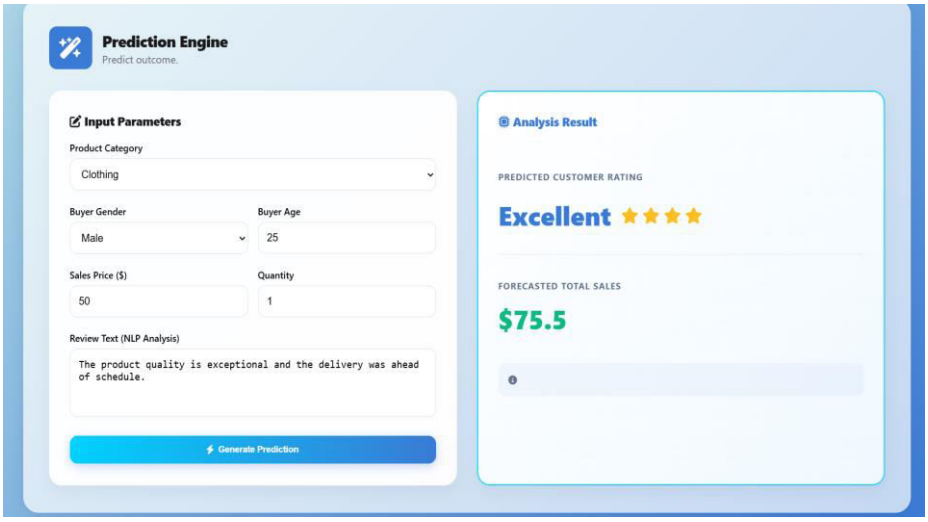

Figure 5: Predictions on test data.

Figure 5 illustrates the prediction interface where users input retail-related parameters and obtain predicted outcomes. The screen displays prediction results generated by the trained ML models in a structured format. This figure demonstrates the successful deployment of the predictive system in a real-time web environment.

4.1 Comparative Analysis

Table. 1: Comparative performance for rating prediction

Model Architecture	Accuracy	Precision	Recall	F1 Score
HS Tree Model	0.8560	0.8802	0.8560	0.8074
TAO Tree Model	0.8529	0.7655	0.8529	0.7981
BRC Model	0.5620	0.5241	0.5620	0.5314
GCN-NGB Model	0.9623	0.9628	0.9623	0.9619

The table 1 leaderboard indicates that the proposed GCN with NGB delivers the best overall performance across all evaluation metrics. It achieves the highest accuracy, precision, recall, and F1-score, demonstrating strong classification capability and consistency. HS Tree and TAO Tree show competitive but comparatively lower performance, with HS Tree performing slightly better in precision. In contrast, the BRC model records the weakest results, indicating limited effectiveness for the classification task.

Table. 2: Comparative performance for sales prediction

Model Architecture	MAE	MSE	RMSE	R^2 Score
HS Tree	33.6095	2564.0703	50.6366	0.8152
TAO Tree	81.2210	13873.8119	117.7871	0.0002
BRC	72.3830	8283.6674	91.0147	0.4031
GCN-NGB Model	26.3185	1274.9229	35.7061	0.9081

In table 2, the proposed GCN model again outperforms all other approaches by achieving the lowest MAE, MSE, and RMSE values along with the highest R^2 score. This highlights its superior predictive accuracy and ability to capture underlying data patterns. HS Tree shows acceptable performance with moderate error levels and a strong R^2 score. However, TAO Tree and BRC suffer from high prediction errors and low R^2 values, reflecting poorer regression performance.

5. CONCLUSION

The research successfully developed and implemented a ML–based retail analytics and prediction system using a web-based framework. The system integrates data preprocessing, EDA, and multiple predictive models to analyze retail transactions and customer behaviour. Models such as HS Tree, TAO Tree, BRC, and the proposed GCN with NGB were implemented and evaluated using real-world retail data. The experimental results demonstrate improved prediction accuracy and stability for the proposed GCN-based model when compared with traditional tree-based approaches. The use of the Flask framework enabled seamless interaction between users and the analytical backend, providing secure authentication, data upload, model execution, and result visualization. The system supports both corporate users for analytics and normal users for prediction, ensuring role-based access and usability. The proposed solution enhances decision-making in retail environments by providing accurate insights into sales performance, customer behaviour, and demand patterns.

REFERENCES

[1] Marketfeed. (2020, October 20). Tata Consumer Products – A detailed analysis. https://www.marketfeed.com/read/en/tata-consumer-products-a-detailed-analysis.

[2] Har, L.L.; Rashid, U.K.; Te Chuan, L.; Sen, S.C.; Xia, L.Y. Revolution of retail industry: From perspective of retail 1.0 to 4.0. Procedia Comput. Sci. 2022, 200, 1615–1625.

[3] Rehman, A.; Naz, S.; Razzak, I. Leveraging big data analytics in healthcare enhancement: Trends, challenges and opportunities. Multimed. Syst. 2022, 28, 1339–1371.

[4] Yusof, Z.B. Analyzing the role of predictive analytics and ML techniques in optimizing inventory management and demand forecasting for e-commerce. Int. J. Appl. Mach. Learn. 2024, 4, 16–31.

[5] Best, J.; Glock, C.H.; Grosse, E.H.; Rekik, Y.; Syntetos, A. On the causes of positive inventory discrepancies in retail stores. Int. J. Phys. Distrib. Logist. Manag. 2022, 52, 414–430.

[6] Suryawanshi, R.; Musale, S.; Bhosale, S. Comparative analysis of use of ML algorithm for prediction of sales. J. Electr. Syst. 2024, 20, 851–863.

[7] Kadam, V.; Vhatkar, S. Design and develop data analysis and forecasting of the sales using ML. In Intelligent Computing and Networking: Proceedings of IC-ICN 2021; Springer: Singapore, 2022; pp. 157–171.

[8] Mahoto, N.A.; Iftikhar, R.; Shaikh, A.; Asiri, Y.; Alghamdi, A.; Rajab, K. An intelligent business model for product price prediction using ML approach. Intell. Autom. Soft Comput. 2021, 30, 1.

[9] Vidhya, V.; Donthu, S.; Veeran, L.; Lakshmi, Y.S.; Yadav, B. The intersection of AI and consumer behavior: Predictive models in modern marketing. Remit. Rev. 2023, 8, 4.

[10] Wan, S.; Chen, J.; Qi, Z.; Gan, W.; Tang, L. Fast RFM model for customer segmentation. In Companion Proceedings of the Web Conference 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 965–972. [Google Scholar]

[11] Bagul, N.; Berad, P.; Surana, P.; Khachane, C. Retail customer churn analysis using RFM model and K-means clustering. Int. J. Eng. Res. Technol. 2021, 10, 3.

[12] Cordova, R.S. Customer segmentation in the online retail industry using big data analytics. J. Theor. Appl. Inf. Technol. 2024, 102, 22.

[13] Aruva, S.P. A Systematic Evaluation of Regressions and Loss Functions for the Prediction of Monetary Value in RFM Analysis. Ph.D. Thesis, National College of Ireland, Dublin, Ireland, 2023.

[14] Rivera-Castro, R.; Pletnev, A.; Pilyugina, P.; Diaz, G.; Nazarov, I.; Zhu, W.; Burnaev, E. Topology-based clusterwise regression for user segmentation and demand forecasting. arXiv 2020, arXiv:2009.03661. Available online: https://arxiv.org/abs/2009.03661 (accessed on 10 December 2024).

[15] Yanchenko, A.K.; Deng, D.D.; Li, J.; Cron, A.J.; West, M. Hierarchical dynamic modeling for individualized Bayesian forecasting. arXiv 2021, arXiv:2101.03408. Available online: https://arxiv.org/abs/2101.03408 (accessed on 10 December 2024).