International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Scalable Retail Demand Forecasting Using XGboost Fusion with Tabular Feature Interpretation

Ch. Mounika[1]*, Mallegari Varshith Reddy[1], Kayithi Rakesh[1], Deshaipet Yashwanth Reddy[1]
[1]Department of Computer Science and Engineering (DS), Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.
*Correspondence: Ch. Mounika

ABSTRACT

Accurate demand forecasting is critical in the retail industry, as it directly influences inventory management, pricing strategies, and overall business profitability. The increasing complexity of customer behavior, seasonal trends, and competitive market dynamics has made traditional forecasting methods less effective. Traditional retail systems primarily relied on manual analysis, historical averages, spreadsheets, or simple statistical methods. These approaches depended heavily on human judgment and past sales trends, offering limited accuracy and scalability. They also lacked automation, real-time prediction capabilities, and the ability to analyze complex relationships among multiple influencing factors, often resulting in overstocking or stock shortages. To address these limitations, the proposed system introduces a machine learning-based solution that automates the entire demand forecasting process. At the core of the system is a hybrid TabNet-Enhanced eXtreme Gradient Boosting (XGBoost) regression model, which leverages the feature learning capabilities of TabNet alongside the predictive power of XGBoost. TabNet effectively extracts important feature representations from tabular retail data, while XGBoost captures non-linear relationships and interactions among features. In addition to the proposed model, baseline models including K-Nearest Neighbors (KNN), Decision Tree Regressor (DTR), and Gradient Boosting Regressor (GBR) are implemented for performance comparison. Experimental results show that the hybrid TabNet-XGBoost model outperforms traditional models, achieving higher accuracy, improved R^2 scores, and lower error metrics. The system is deployed through a Flask-based interactive desktop interface, providing role-based access that allows AIML Engineers to perform model training, exploratory data analysis, and performance evaluation, while Retailers can perform real-time demand and discount predictions.

Keywords: Demand forecasting, Retail analytics, Inventory management, Pricing strategy, Feature extraction, Nonlinear data analysis, Predictive systems, Sales prediction, Seasonal trends, Customer behavior analysis.

1. INTRODUCTION

The retail industry has become highly informative through the collection of large amounts of transactional data every day [1]. 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 [2]. 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 as demonstrate in Fig. 1, which requires accurate demand forecasting and consumer segmentation as critical success factors [3]. 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 [4]. With changing business dynamics, where most decisions are made based on data, the use of machine

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

learning, statistical modeling for data analysis, and advanced segmentation techniques have become a solution to these problems.

Fig. 1: Retail Sales Prediction Demand Forecasting.

2. LITERATURE SURVEY

Shirole et al. [5] presented the system design that combines domain knowledge for improving EDA process using the VizML framework, based on guided analytics. The approach entails capturing EDA sessions of the domain experts through the storing of their interactions and context into the interaction and context storage system. These stored interactions are then used to suggest sequences of analysis steps for domain newbies, who are always in one way the direct consumer of the findings performing in a similar dataset and guiding them to useful discoveries. Rajan et al. [6], employed IoT analytics and marketing intelligence with a view to facilitating decision-making within the complex digital context. The authors presented a method involving data acquisition utilizing the IoT devices, data gathering, and data pre-processing, followed using machine learning-based client-side analytics including clustering, classification, and regression. Bibliographically, retail demand forecasting studies predominantly employ univariate statistical techniques such as ARIMA or traditional machine learning techniques, namely tree-boosting regressors [7].

Efat et al. [8] proposed a hybrid sales forecasting model that combines adaptive trend estimated series (ATES) with a deep neural network based on a LSTM architecture to capture dynamic and nonlinear sales patterns at the product-specific store level. Liu et al. [9] proposed a hybrid electric vehicle sales forecasting method combining ML-based sentiment analysis and secondary decomposition. Haque et al. [10] improved retail selling forecast by adding macroeconomic indicators, such as CPI, ICS, and unemployment rates, into a dataset that embraces selling records from five years ago collected from Walmart. Using Lasso, Ridge Regression models, LightGBM, XGBM, and Decision Trees, this paper assesses the impact of these macroeconomic predictors on prediction bias. When macroeconomic factors are included, it is established that there are slight but significant enhancements achieved in model performance enhancement, whereby the LightGBM model is found to enhance the best value of the RMSE of 1.715 and MAE of 0.847.

Kasem et al [11] addressed these gaps by combining regression-based forecasting with recency, frequency, monetary (RFM)-based segmentation to offer a more comprehensive analytical approach.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Naik et al. [12] outlines a strategic plan to establish a reliable Customer Segregation Infrastructure with the help of data mining. The method included data acquisition from multiple organizational sources and data preprocessing, exploratory data analysis, and feature selection for data relevance. Clustering, classification, and association mining data mining algorithms were then used, to uncover underlying patterns for accurate customer segmentation.Cao et al. [13] Recent advancements in deep learning have demonstrated the efficacy of models like Long Short-Term Memory (LSTM) networks and Transformer-based architectures in time-series forecasting. For instance, a study published in Scientific Reports proposes a time series prediction model that fuses Transformer and LSTM algorithms, highlighting the strengths of both approaches in capturing temporal dependencies.

Langer et al. [14], researched IoT analytics and marketing intelligence with a view to facilitating decision-making within the complex digital context. The authors presented a method involving data acquisition utilizing the IoT devices, data gathering, and data pre-processing, followed using machine learning-based client-side analytics including clustering, classification, and regression. Lewaaelhamd et al. [15], used the RFM (recency, frequency, monetary) model with the help of the K-means clustering technique to classify customers according to their buying habits. The researchers use data obtained from an e-commerce platform in the UK that included transactions that occurred between the years 2010 and 2011 and clean the data by deleting any incomplete information.

3. PROPOSED SYSTEM

The proposed system introduces an intelligent and automated approach for retail sales demand and discount prediction using advanced machine learning techniques. The system is designed to overcome the limitations of traditional manual forecasting methods by integrating data preprocessing, exploratory analysis, model training, evaluation, and prediction into a single unified pipeline as demonstrated in Fig. 2. The complete workflow of the proposed system is explained step by step as follows

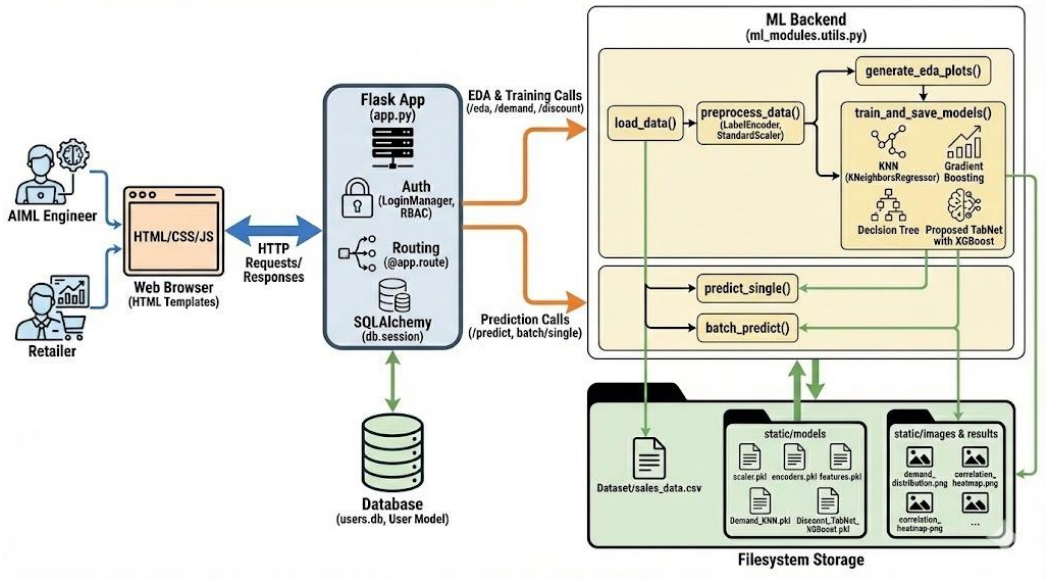


Fig. 2: Proposed System Architecture.

Data Collection: The system begins by collecting historical retail sales data from a structured dataset stored in CSV format. The dataset includes various features such as product category, region, seasonality, inventory level, units sold, competitor value, and pricing-related attributes, which influence demand and discount decisions.

Data Preprocessing: The collected data undergoes preprocessing to ensure quality and consistency. Missing numerical values are handled using mean imputation, while missing categorical values are replaced with the most frequent category. Categorical attributes are converted into numerical format

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

using label encoding, and numerical features are standardized using a scaling technique. The preprocessing artifacts are saved to maintain consistency during prediction.

Data Analysis: Exploratory Data Analysis is performed to understand data distribution and relationships between variables. The system generates visualizations such as histograms, box plots, correlation heatmaps, and scatter plots to analyze demand patterns, discount behavior, seasonal trends, and category-wise performance.

Feature Selection and Target Identification: Relevant features are selected based on the prediction objective. For demand prediction, discount-related attributes are included as features, whereas for discount prediction, demand is excluded to avoid data leakage. This ensures accurate and unbiased model training.

Model Training: Multiple machine learning regression models are trained, including KNN, DTR, GBR, and the proposed TabNet-XGBoost model. The hybrid model combines feature learning capabilities of TabNet with the predictive strength of XGBoost to capture complex retail data patterns.

Model Evaluation: The trained models are evaluated using standard performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score. Actual versus predicted graphs are generated to visually assess model performance and accuracy.

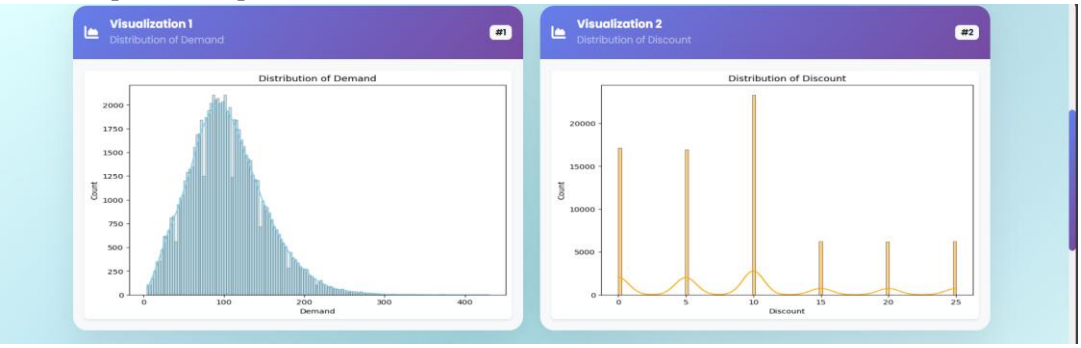
Model Saving and Deployment: The best-performing models, along with encoders and scalers, are saved for reuse. These trained models are deployed within an interactive Python-based interface, enabling real-time predictions without retraining.

Prediction and User Interaction: The system provides prediction functionalities through a Flask-based interface. Retailers can perform single input prediction, random sample analysis, or batch prediction using uploaded datasets. Role-based access ensures secure usage for AIML Engineers and Retailers.

Decision Support: Finally, the predicted demand and discount values support informed decision-making related to inventory planning, pricing strategies, and sales optimization.

4. RESULTS ANALYSIS

This section describes the results obtained from the implementation of the retail demand forecasting system. The figures presented illustrate the complete workflow of the system, starting from user interaction and dataset management to exploratory data analysis, model evaluation, and final demand prediction. Each figure represents a key stage in the functioning of the proposed machine learning–based demand prediction platform.



International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

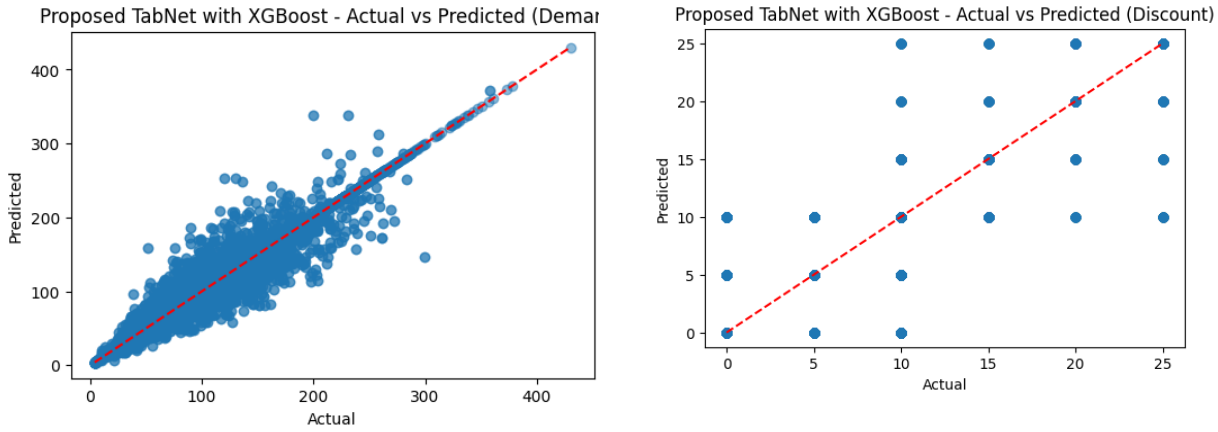
Fig. 3: Exploratory data analysis on demand forecasting data. (a) Histogram. (b) Histogram. (c) Boxplot. (d) Heatmap. (e) Scatter plot. (f) Pie chart.

Fig. 3 illustrates the results of the exploratory data analysis conducted on the retail sales dataset. These visualizations provide insights into the distribution, relationships, and patterns in key features, supporting feature selection, preprocessing, and model design for demand and discount prediction.

- Fig. 3(a) Distribution of Demand: A histogram with KDE overlay is generated using sns.histplot() to show the frequency of product demand values. The visualization highlights that most products exhibit mid-range demand, while a few products have extremely high or low demand. This information guides scaling and model sensitivity to extreme values.
- Fig. 3(b) Distribution of Discount: A histogram with KDE overlay illustrates the spread of discounts across products. Most discounts are concentrated in lower to mid ranges, with very few extreme discounts. The plot helps identify typical promotional patterns that influence the discount prediction model.
- Fig. 3(c) Demand by Category: A boxplot created using sns.boxplot() compares demand distributions across product categories. It shows which categories consistently achieve higher sales, which have moderate sales, and which contain outlier products with unusually high or low demand. This assists in understanding category-based demand variation for feature engineering.
- Fig. 3(d) Correlation Heatmap: A heatmap generated using sns.heatmap() visualizes correlations among numeric features such as Demand, Discount, Units Sold, and Inventory Level. Strong positive correlations, for example between Units Sold and Demand, indicate key relationships that affect sales performance. The heatmap helps identify influential features for predictive modeling.
- Fig. 3(e) Units Sold vs Demand: A scatter plot using sns.scatterplot() depicts the relationship between units sold and product demand. The plot shows a general upward trend, confirming that higher demand results in more units sold, while the spread of points highlights variability, which supports the use of models capable of capturing non-linear trends.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

- Fig. 3(f) Region-wise Demand Share: A pie chart represents the proportion of total demand contributed by each region. Some regions dominate total demand while others contribute less, highlighting geographic concentration in sales. This insight informs regional inventory management and targeted marketing strategies.



(a) (b)

Fig. 4: Scatter Plots of target attributes from TabNet-XGBoost regressor model. (a) Demand, (b) Discount.

Fig. 4 the scatter plots depict the performance of the proposed hybrid TabNet-XGBoost model, combining TabNet's feature representation learning with XGBoost's gradient boosting for highly accurate regression. TabNet extracts latent feature embeddings capturing complex interactions among variables, while XGBoost refines predictions iteratively.

- **(a) Demand:** Predicted demand points align tightly with actual values across all ranges, including low, mid, and high demand products. The model captures intricate relationships between features such as units sold, seasonality, region, and inventory levels. Residuals are minimal, demonstrating superior generalization and the ability to learn subtle patterns in the dataset.
- **(b) Discount:** Discount predictions show a high level of precision with points clustered closely along the diagonal. The model accurately reflects promotional strategies, category-specific discounts, and the influence of competitor pricing. Prediction consistency across the full discount range highlights the model's robustness for practical retail decision-making.

Demand Prediction - Model Performance

Model	MAE	MSE	RMSE	R2 Score
KNN	0.1358	3.3061	0.0182	0.8476
Gradient Boosting	0.2052	7.3638	0.0271	0.6606
Decision Tree	0.3121	16.0294	0.04	0.2612
Proposed TabNet with XGBoost	0.0275	0.8964	0.0095	0.9587

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

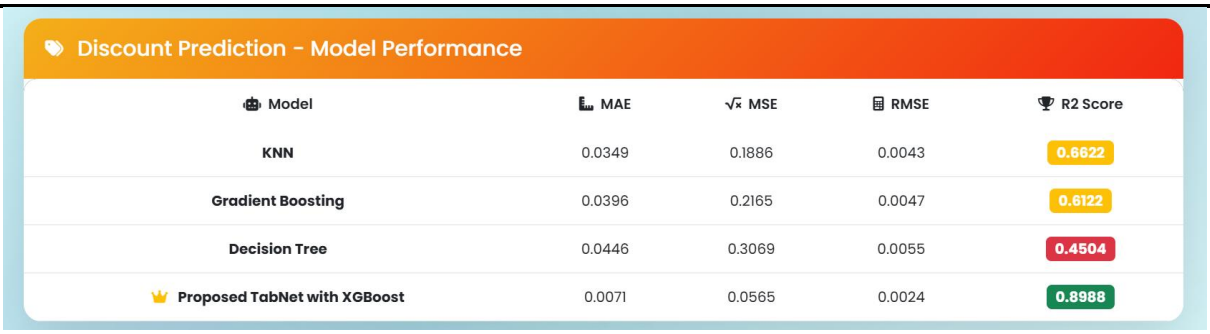


Fig. 5: Models performance comparison

Fig. 5 provides a comparative analysis of all implemented models, including KNN, DTR, GBR, and TabNet-XGBoost. The comparison highlights evaluation metrics such as error rates and prediction accuracy. The figure clearly shows that the TabNet-XGBoost model achieves the best overall performance. This comparison validates the effectiveness of the proposed hybrid approach for retail demand forecasting.

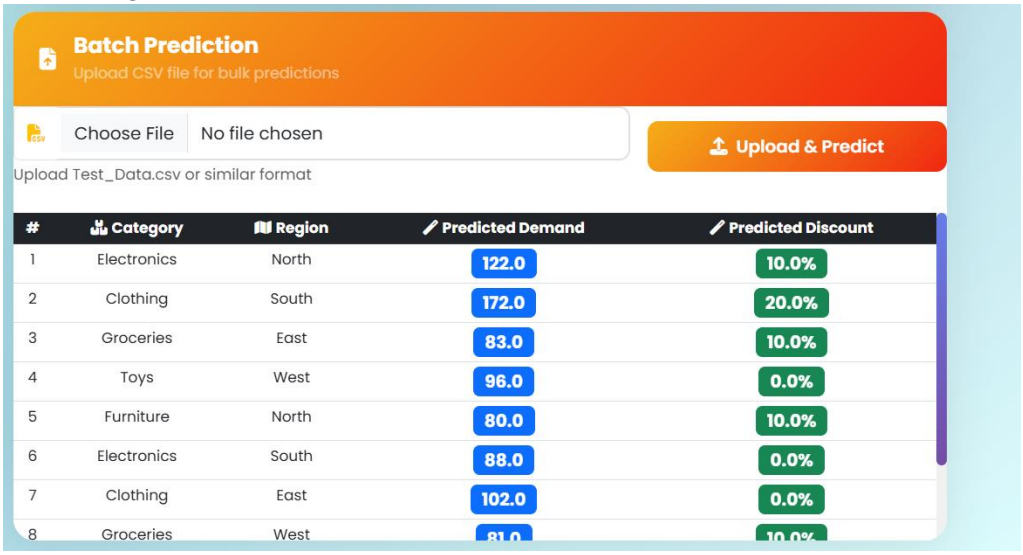


Fig. 6: Presents the predictions screen

Fig. 6 illustrates the prediction screen, where users input retail parameters such as category, region, inventory level, discount, promotion status, weather condition, and seasonality. The trained TabNet-XGBoost model processes these inputs and generates accurate demand predictions. This screen represents the final outcome of the system, delivering actionable insights that assist in inventory planning and decision-making.

Table 1: Demand prediction model performance

Model	MAE	MSE	RMSE	R^2 Score
KNN Model	0.1358	3.3061	0.0182	0.8476
GBR Model	0.2052	7.3638	0.0271	0.6606
DTR Model	0.3121	16.0294	0.0400	0.2612

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Proposed TabNet-XGBoost	**0.0275**	**0.8964**	**0.0095**	**0.9587**

Table 1 presents the performance metrics of four regression models used for predicting product demand. Among the baseline models, the KNN Regressor achieves moderate accuracy with an R^2 score of 0.8476, indicating reasonable alignment between predicted and actual demand. The GBR performs slightly lower with an R^2 of 0.6606, while the DTR shows limited predictive capability, reflected by its low R^2 of 0.2612 and higher error values. The proposed TabNet-XGBoost hybrid model outperforms all baseline models, achieving the highest R^2 score of 0.9587 along with the lowest MAE, MSE, and RMSE, demonstrating superior accuracy and robustness in capturing complex relationships within the retail dataset. These results confirm the effectiveness of the hybrid approach for precise demand forecasting.

Table 2: Discount Prediction – Model performance

Model	MAE	MSE	RMSE	R^2 Score
KNN Model	0.0349	0.1886	0.0043	0.6622
GBR Model	0.0396	0.2165	0.0047	0.6122
DTR Model	0.0446	0.3069	0.0055	0.4504
Proposed TabNet-XGBoost	**0.0071**	**0.0565**	**0.0024**	**0.8988**

Table 2 summarizes the performance of four regression models for predicting product discounts. Among the baseline models, the KNN Regressor shows moderate accuracy with an R^2 score of 0.6622, while the GBR performs slightly lower with an R^2 of 0.6122. The DTR exhibits limited predictive capability, reflected in its lower R^2 of 0.4504 and higher error metrics. The proposed TabNet-XGBoost model outperforms all baseline models, achieving the highest R^2 score of 0.8988 along with the lowest MAE, MSE, and RMSE, demonstrating its ability to accurately capture complex feature interactions and provide precise discount predictions.

5. CONCLUSION

This research successfully designed and implemented an intelligent Retail Sales Demand Prediction System using machine learning and deep learning techniques. The system integrates data preprocessing, exploratory data analysis, model training, evaluation, and prediction into a unified web-based platform. By utilizing real-world retail attributes such as inventory level, units sold, discounts, promotions, weather conditions, seasonality, and competitor pricing, the system effectively captures complex demand patterns. Multiple predictive models including KNN Regressor, DTR, and GBR were implemented and evaluated. Among these, the proposed TabNet-XGBoost hybrid model achieved superior performance due to its ability to perform adaptive feature selection and ensemble-based learning. The hybrid approach enhanced prediction accuracy and stability while handling both numerical and categorical features efficiently. The integration of secure admin and user modules ensured proper dataset management, model training, and real-time prediction. The results demonstrate that the proposed system provides accurate demand forecasts, reduces dependency on manual analysis,

and supports data-driven decision-making. This research proves that advanced machine learning and deep learning models significantly improve retail demand forecasting and inventory planning processes.

REFERENCES

[1] 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.

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

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

[4] 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.

[5] Shirole, R.; Salokhe, L.; Jadhav, S. Customer segmentation using RFM model and K-means clustering. Int. J. Sci. Res. Sci. Technol. 2021, 8, 591–597.

[6] Rajan, P. Integrating IoT analytics into marketing decision-making: A smart data-driven approach. Int. J. Data Inf. Intell. Comput. 2024, 3, 12–22.

[7] DataRK1. Customers Clustering: K-Means, DBSCAN, and Affinity Propagation. Kaggle. 2023. Available online: https://www.kaggle.com/code/datark1/customers-clustering-k-means-dbscan-and-ap (accessed on 1 January 2025).

[8] Efat, M.I.A.; Hajek, P.; Abedin, M.Z.; Azad, R.U.; Jaber, M.A.; Aditya, S.; Hassan, M.K. Deep-learning model using hybrid adaptive trend estimated series for modelling and forecasting sales. Ann. Oper. Res. 2024, 339, 297–328.

[9] Liu, J.; Pan, H.; Luo, R.; Chen, H.; Tao, Z.; Wu, Z. An electric vehicle sales hybrid forecasting method based on improved sentiment analysis model and secondary decomposition. Eng. Appl. Artif. Intell. 2025, 150, 110561.

[10] Haque, M.S.; Amin, M.S.; Miah, J. Retail demand forecasting: A comparative study for multivariate time series. arXiv 2023, arXiv:2308.11939.

[11] Kasem, M.S.; Hamada, M.; Taj-Eddin, I. Customer profiling, segmentation, and sales prediction using AI in direct marketing. arXiv 2023, arXiv:2302.01786. Available online: https://arxiv.org/abs/2302.01786 (accessed on 1 January 2025).

[12] Naik, S. Customer segregation infrastructure: Unveiling insights through data mining. Int. J. Artif. Intell. 2023, 3, 1–4.

[13] Cao, K.; Zhang, T.; Huang, J. Advanced hybrid LSTM-transformer architecture for real-time multi-task prediction in engineering systems. Sci. Rep. 2024, 14, 4890.

[14] Langer, T.; Meisen, T. System design to utilize domain expertise for visual exploratory data analysis. Information 2021, 12, 140.

[15] Lewaaelhamd, I. Customer segmentation using machine learning model: An application of RFM analysis. J. Data Sci. Intell. Syst. 2024, 2, 29–36.