JAAFR INTERNATIONAL RESEARCH JOURNAL
JOURNAL OF ADVANCE AND FUTURE RESEARCH
JAAFR.ORG | ISSN : 2984-889X
An International Open Access, Peer-reviewed, Refereed Journal

PRODUCT DEMAND FORECASTING WITH CONSUMER BEHAVIOUR CLUSTERING

Madhu Bhashini . S
III BCA STUDENT
Department of Computer Applications(UG),
School of Computing Sciences,
VISTAS, Chennai.
madhushankar110@gmail.com
Dr. V. Divya
Assistant Professor
Department of Computer Applications(UG),
School of Computing Sciences,
VISTAS, Chennai.
divyavenkatraman1992@gmail.com

ABSTRACT

Accurate demand forecasting and consumer behaviour analysis are vital for optimizing retail operations, yet traditional methods struggle with complex, non-linear relationships among pricing, promotions, seasonality, and external factors. This paper presents an integrated data-driven framework combining demand forecasting with consumer behaviour clustering.

The system applies ensemble models — Random Forest, Gradient Boosting, XGBoost, and LightGBM — for robust prediction, alongside K-Means clustering to segment consumers into groups such as price-sensitive, high-value, and seasonal buyers. Feature engineering and temporal transformations further enhance model performance.

A Streamlit-based analytics platform enables real-time visualization, model evaluation, and user-driven demand prediction. Results show ensemble methods outperform individual models, while clustering yields actionable behavioural insights.

By unifying forecasting and segmentation, the framework supports smarter inventory, pricing, and marketing decisions. It is scalable, efficient, and extensible toward deep learning and real-time data integration.

KEYWORDS *— Demand Forecasting, Consumer Behaviour Clustering, Machine Learning, Ensemble Learning, Retail Analytics*

I. INTRODUCTION

In today's dynamic retail landscape, accurate demand forecasting and consumer behaviour analysis are essential for sustained growth. Businesses face constant demand fluctuations driven by pricing, promotions, seasonality, and external factors — and poor forecasting leads to overstocking, stockouts, rising costs, and reduced customer satisfaction.

Traditional statistical and time-series models often fail to capture the complex, non-linear relationships in large retail datasets, and struggle to adapt to fast-changing market conditions. Machine learning has emerged as a stronger alternative, offering greater accuracy and scalability across multiple influencing factors.

Consumer behaviour analysis is equally critical. Customers vary in price sensitivity, preferences, and seasonal purchasing patterns, and segmenting them through techniques like K-Means clustering enables targeted marketing and personalized engagement.

This paper proposes an integrated framework combining ensemble-based demand forecasting with consumer behaviour clustering, supported by an interactive analytics dashboard for real-time visualization and user-driven predictions. The unified system aims to enhance inventory management, pricing optimization, and customer relationship management — delivering an intelligent, adaptable solution for modern retail decision-making.

II. RELATED WORK

A) Demand Forecasting Using Statistical and Machine Learning Models

Demand forecasting has traditionally relied on statistical time-series models such as ARIMA and exponential smoothing. Hyndman and Athanasopoulos [1] demonstrated that these models are effective in capturing seasonality and trends in structured retail data. However, such approaches are limited in handling non-linear relationships and high-dimensional datasets.

To overcome these limitations, machine learning models have been widely adopted. Breiman [2] introduced Random Forest, which improves prediction accuracy through ensemble decision trees. Friedman [3] proposed Gradient Boosting, which iteratively reduces prediction errors. More recently, Chen and Guestrin [4] developed XGBoost, a scalable boosting framework that achieves high performance in large datasets, while Ke et al. [5] introduced LightGBM for faster and more efficient training. These models have significantly improved forecasting accuracy in Retail applications.

B) Ensemble Learning for Demand Prediction

Ensemble learning techniques combine multiple models to improve prediction robustness and accuracy. Studies have shown that ensemble approaches such as voting and stacking reduce both bias and variance in forecasting tasks [6]. Makridakis et al. [7] conducted a large-scale evaluation of forecasting methods and found that hybrid and ensemble approaches consistently outperform individual statistical and machine learning models. These findings support the use of ensemble techniques in complex demand forecasting scenarios.

C) Consumer Behaviour Clustering Techniques

Consumer behaviour analysis is essential for understanding purchasing patterns in retail systems. MacQueen [8] introduced the K-Means clustering algorithm, which is widely used for segmenting customers based on similarity in behaviour. Rokach and Maimon [9] demonstrated that clustering techniques can effectively classify customers into meaningful groups such as high-value, price-sensitive, and low-frequency buyers.

Clustering enables businesses to design targeted marketing strategies and improve customer engagement. However, most existing approaches treat clustering as a standalone analytical tool, without integrating it directly into forecasting systems.

D) Integration of Clustering and Forecasting

Recent research has explored combining clustering with forecasting models to improve prediction performance. Cluster-based forecasting approaches group similar data points before applying predictive models, which helps reduce variability and improves accuracy [10]. Studies also indicate that integrating clustering with machine learning models enhances generalization and enables better handling of heterogeneous data [11].

Despite these advancements, many existing methods lack a unified framework that combines clustering, forecasting, and real-time analytics in a single system.

E) Interactive Analytics and Visualization Systems

The development of interactive dashboards has improved the usability of data-driven systems. Tools such as Streamlit enable rapid deployment of machine learning models with real-time user interaction [12]. Géron [13] emphasizes that visualization plays a crucial role in interpreting model outputs and supporting decision-making.

However, most existing systems focus primarily on either predictive modeling or visualization, with limited integration of consumer behaviour analysis and real-time forecasting capabilities.

F) Research Gap

From the literature, it is evident that demand forecasting, ensemble learning, and consumer behaviour clustering have been extensively studied. However, these components are often treated independently. Limited research has focused on integrating ensemble-based demand forecasting with consumer behaviour clustering within an interactive real-time analytics framework.The proposed work addresses this gap by developing a unified system that combines these components, enabling more accurate predictions and improved decision-making in retail environments.

III. SYSTEM ARCHITECTURE

A) Overview

The proposed system is an integrated framework for product demand forecasting and consumer behaviour clustering. It follows a structured pipeline that transforms raw retail data into actionable insights and real-time predictions. The workflow is defined as:

Dataset → Preprocessing → Feature Engineering → Clustering → Forecasting → Dashboard → Prediction

B) Data Processing and Feature Engineering

The system utilizes a retail dataset containing product, pricing, sales, and inventory information. Data preprocessing includes handling missing values, removing inconsistencies, and formatting temporal attributes.

Feature engineering is applied to extract meaningful variables such as time-based features (month, day), pricing metrics (discounted price, price difference), and inventory indicators (stock utilization), which improve model performance.

C) Consumer Behaviour Clustering

Consumer segmentation is performed using the **K-Means clustering algorithm**. Relevant behavioural features are normalized and grouped into clusters representing different customer segments such as high-value and price-sensitive consumers. This step provides insights into purchasing patterns and supports targeted decision-making.

D) Demand Forecasting

The system employs **ensemble machine learning models**, including Random Forest, XGBoost, and LightGBM, to predict product demand. Ensemble techniques such as voting and stacking are used to improve prediction accuracy and robustness.

E) Visualization and Real-Time Prediction

An interactive dashboard is developed using **Streamlit** with **Plotly-based visualizations**. It enables users to explore data, analyze model performance, and generate real-time demand predictions by providing input parameters such as price, discount, and inventory level.

F) Summary

The architecture integrates clustering, forecasting, and visualization into a unified system, enabling accurate predictions and effective decision-making in retail environments.

IV. IMPLEMENTATION

A) Tools and Technologies

The system is developed using Python, leveraging libraries such as Scikit-learn for preprocessing and model building. Advanced boosting models are implemented using XGBoost and LightGBM. Data visualization is handled using Plotly.

B) Processing Workflow

A structured pipeline is implemented to handle data preprocessing, feature generation, clustering, and model training. This ensures smooth data flow and consistency across different stages of the system.

C) Clustering Module

The clustering component is implemented using K-Means, where normalized input data is grouped into distinct segments. The output clusters are used to analyze behavioural trends and support better interpretation of demand patterns.

D) Forecasting Module

The forecasting component includes multiple trained models, and their predictions are combined using ensemble methods. The final selected model is integrated into the system to generate demand forecasts.

E) User Interface and Visualization

A web-based interface is created using Streamlit to allow users to interact with the system. It provides features for data visualization, cluster analysis, and model comparison. Interactive graphs are generated using Plotly for better understanding of trends.

F) Real-Time Prediction

The system enables users to input parameters such as pricing, discount, and inventory levels. Based on these inputs, the trained model produces real-time demand predictions. This functionality supports quick and informed decision-making.

V. EVALUATION

A) Forecasting Model Performance

The performance of the demand forecasting models was evaluated using standard regression metrics, including RMSE, MAE, R^2, and MAPE. Multiple models were trained and compared to identify the most accurate approach.

Table 1. Model Performance Comparison

MODEL	RMSE	MAE	R^2 Score
Random Forest	18.5	12.3	0.89
Gradient Boosting	16.8	11.1	0.91
XGBoost	15.2	10.4	0.93
LightGBM	14.7	10.1	0.94

The results indicate that LightGBM achieved the best performance, with the lowest RMSE and highest R^2 score. Ensemble methods improved prediction accuracy by effectively capturing non-linear relationships in the data.

B) Ensemble Model Evaluation

To further enhance performance, ensemble techniques such as voting and stacking were applied. The combined model demonstrated improved stability and reduced prediction error compared to individual models**.**

Table 2. Ensemble Model Performance

MODEL TYPE	RMSE	MAE	R^2 Score
Individual Best Model	14.7	10.1	0.94
Voting Ensemble	13.9	9.5	0.95

Stacking Ensemble	13.2	9.1	0.96

The stacking ensemble achieved the best overall performance, confirming that combining multiple models enhances prediction robustness.

C) Clustering Evaluation

The effectiveness of consumer behaviour clustering was evaluated using the Elbow Method and Silhouette Score. The optimal number of clusters was found to be k = 3, representing distinct consumer segments**.**

Table 3. Clustering Evaluation

METRIC	VALUE
Optimal clusters (k)	3
Silhouette Score	0.64

The clustering results indicate well-separated groups, enabling meaningful segmentation such as high-value, moderate, and price-sensitive customers.

D) System Performance

The responsiveness of the system was evaluated based on prediction time and dashboard interaction. The average time required to generate a demand prediction was less than 1 second, ensuring real-time usability.

The Streamlit dashboard demonstrated smooth interaction with minimal latency, allowing users to visualize data and obtain predictions efficiently.

E) Comparative Analysis

The proposed system was compared with traditional statistical models to evaluate performance improvements.

Table 4. Comparison with Traditional Methods

METHOD	RMSE	ACCURACY LEVEL
ARIMA	21.5	Moderate
Random Forest	13.9	Good
Proposed Ensemble	13.2	High

The results show that the proposed system significantly outperforms traditional methods, highlighting the effectiveness of ensemble learning and integrated clustering.

F) summary

The evaluation demonstrates that the proposed system achieves high prediction accuracy, effective customer segmentation, and real-time performance. The integration of ensemble learning with clustering provides a reliable and scalable solution for retail demand forecasting.

VI. DISCUSSION

A) Strengths

The proposed system demonstrates several key strengths. First, the use of ensemble machine learning models improves demand prediction accuracy compared to individual models. By combining algorithms such as Random Forest, XGBoost, and LightGBM, the system effectively captures complex and non-linear relationships in retail data.

Second, the integration of consumer behaviour clustering provides deeper insights into purchasing patterns. Segmenting customers into meaningful groups enables more targeted decision-making in areas such as pricing and marketing strategies.

Additionally, the inclusion of an interactive dashboard enhances usability by allowing real-time data visualization and prediction. This makes the system practical and accessible for non-technical users.

B) Limitations

Despite its advantages, the system has certain limitations. The performance of the forecasting models depends heavily on the quality and size of the dataset. Incomplete or noisy data can negatively affect prediction accuracy.

The K-Means clustering algorithm assumes spherical clusters and requires predefinition of the number of clusters, which may not always represent real-world consumer behaviour accurately.

Furthermore, the current system primarily focuses on historical data and does not incorporate real-time external factors such as market trends, economic conditions, or sudden demand shifts.

C) Future Work

Future enhancements can further improve the system's performance and applicability. Deep learning models such as LSTM can be incorporated to better capture temporal dependencies in demand data.

The system can also be extended by integrating real-time data sources, such as APIs, to enable dynamic forecasting. Advanced clustering techniques, including DBSCAN or hierarchical clustering, can be explored for improved segmentation.

Additionally, the development of a personalized recommendation system based on consumer behaviour could further enhance customer engagement and business outcomes.

VII. CONCLUSION

This paper presented an integrated framework for product demand forecasting combined with consumer behaviour clustering to support data-driven decision-making in retail environments. The proposed system leverages ensemble-based machine learning models, including Random Forest, XGBoost, and LightGBM, to improve prediction accuracy, while K-Means clustering is employed to segment consumers based on purchasing patterns.

The experimental results demonstrate that ensemble techniques outperform individual models in terms of accuracy and robustness. In addition, the clustering approach provides meaningful insights into customer behaviour, enabling better understanding of demand patterns. The integration of these components within an interactive Streamlit-based dashboard allows real-time visualization and prediction, enhancing the practical usability of the system.

Overall, the proposed framework offers a scalable and efficient solution for retail analytics, contributing to improved inventory management, optimized pricing strategies, and enhanced customer engagement. The study highlights the effectiveness of combining predictive modeling with behavioural analysis in a unified system.

Future improvements may include the incorporation of deep learning techniques for time-series forecasting, integration of real-time data sources, and development of advanced recommendation systems to further enhance system capabilities.

REFERENCES

[1] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Melbourne, Australia: OTexts, 2018.

[2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[5] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

[6] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward," *PLOS ONE*, vol. 13, no. 3, 2018.

[7] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967, pp. 281–297.

[8] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*. Singapore: World Scientific, 2008.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[10] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.

[12] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] Streamlit Inc., "Streamlit: The fastest way to build data apps," 2023. [Online]. Available: https://streamlit.io