Journal of Theoretical and Applied Information Technology
15th May 2024. Vol.102. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EVALUATING THE PERFORMANCE OF XGBOOST AND GRADIENT BOOST MODELS WITH FEATURE EXTRACTION IN FMCG DEMAND FORECASTING: A FEATURE-ENRICHED COMPARATIVE STUDY

MURARI THEJOVATHI[1], DR M.V.P. CHANDRA SEKHARA RAO[2]

[1]Department of Computer Science and Engineering, Acharya Nagarjuna University Guntur Andhra Pradesh, India
[2]Department of Computer Science and Engineering, RVR&JC College of Engineering, Guntur, Andhra Pradesh, India
E-mail: [1] kkutheju@gmail.com , [2] manukondach@gmail.com

ABSTRACT

In this paper We are proposing the inclusion of Gradient Boost, another ensemble technique, to broaden the scope and potentially improve forecasting accuracy. This research looks at how XGBoost and Gradient Boost, two powerful ensemble learning methods, can be used to predict demand in the FMCG sector. The suggested method also includes advanced feature extraction techniques to make the model work better. The current method uses XGBoost, a well-known and effective gradient-boosting technique that is fast and easy to scale. The suggested method includes gradient boost, which is another ensemble technique, as well as feature extraction techniques that help find and use the dataset's most important information. The research aims to compare the performance of XGBoost and Gradient Boost models in the context of demand forecasting for Fast-Moving Consumer Goods (FMCG) data. Additionally, the study incorporates feature extraction methods to enhance the models' predictive capabilities. We test both models thoroughly using FMCG data to see how well they work in terms of accuracy, reliability, and how quickly they can be run. To find the factors that have the most influence on demand prediction, feature extraction techniques like Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) are used. The study's results tell us a lot about how well the XGBoost and Gradient Boost models work for predicting demand in the FMCG sector. Using feature extraction methods is also meant to find hidden patterns in the data, which will help supply chain professionals in the FMCG business make more accurate predictions and better decisions. Researchers can use the results of this study to help them choose the best method for their own demand forecasting needs. This will improve operational efficiency and cut costs in the FMCG supply chain.

Keywords: *GaradientBoost, XGBoost, FMCG Sector, Feature Extraction, Demand forecasting*

1. INTRODUCTION

Demand forecasting is an important part of supply chain management in the Fast-Moving Consumer Goods (FMCG) business because it helps keep inventory levels low and makes sure that customers can get the products they want. XGBoost, also known as eXtreme Gradient Boosting, represents a strong and scalable implementation of the gradient boosting framework. It is highly proficient in managing structured and tabular data and is extensively utilized for a range of machine-learning assignments. Highlighted features consist of regularization techniques, parallel computing, and effective management of missing values. XGBoost has become popular because of its efficiency, precision, and capability to manage extensive datasets.

Gradient boosting is an ensemble learning technique that builds a predictive model by incrementally including weak learners, often decision trees. It improves a loss function via gradient descent, with an emphasis on correcting deficiencies from previous models. Gradient boosting is known for its robust predictive power and adaptability, making it valuable for regression and classification tasks. Feature extraction is an essential process in machine learning where useful features are chosen from raw data to enhance model performance. It aids in reducing dimensionality, improving interpretability, and alleviating the curse of dimensionality. By extracting significant features, the model may prioritize pertinent information, resulting in improved generalization and enhanced efficiency. Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) are frequently used techniques for feature extraction.

Paper Highlights

- Analyze FMCG sales data using statistical and machine learning techniques to identify consumer demand trends in urban and rural areas.
- Develop forecasting models like XGBoost and Gradient Boost model, ensuring data pre-processing for accuracy. And Evaluate models' effectiveness through performance metrics such as accuracy, precision, and recall, reliability, and computational efficiency across different scenarios.
- Compare the performance of XGBoost and Gradient Boost models specifically in the context of demand forecasting for FMCG data. Investigates the use of Introduces advanced feature extraction techniques to enhance model performance, aiming to identify and utilize the most important information in the dataset.
- Proposes the inclusion of Gradient Boost, another ensemble technique, and emphasizes the use of feature extraction techniques to improve the model's performance.
- Highlights the significance of the study's results for supply chain professionals in the FMCG business. The improved accuracy and hidden pattern discovery are expected to assist in making more accurate predictions and better decisions.

2. LITERATURE SURVEY

The study discovered that Gradient Boosted Decision Trees (GBDT) with Feature Interaction improve customer predicting for organizations. These strategies capture complicated variable interactions to improve demand estimations.

The paper's multidimensional demand forecasting method incorporates historical time-series data, product attributes, customer preferences, and environmental influences like weather and the economy. This holistic approach lets businesses construct models that go beyond prediction to comprehend high-demand item dynamics in different situations.

The study demonstrates demand forecasting affects more than retail. Demand forecasting affects inventory management, personnel scheduling, and supply chain optimization, making it a critical tool for operational efficiency.

Better forecasting and consumer-driven operations are recommended in the research. In the competitive retail industry, these components must work together to improve customer service and profitability. The new paradigm drives companies to prioritize customers and compete with current technology.

The study emphasizes the need of precise demand forecasting for enterprises to accommodate customer preferences and make educated operational decisions. According to the paper, demand forecasting helps companies adapt swiftly to market and consumer developments.

GBDT with Feature Interaction is applied in rural and urban environments, a highlight. This adaptability shows the technology's potential for global precision forecasting.

In the dynamic retail industry, advanced demand forecasting methods are necessary for economic sustainability, operational optimization, and data-driven decision-making. The paper's comprehensive approach, which analyses various variables and strategic concerns, makes demand forecasting essential to modern firm strategy. Modern technologies and methods are vital, suggesting retail innovation and adaptation.

Researcher(s)	***Methodology Employed***	***Evaluation Criteria***	***Data Utilized***	***Model Accuracy***
P. M. Pardalos, R. J. Hyndman, Y. Khandakar	Time Series Forecasting, forecast Package for R	Error Metrics: MAE, MSE, RMSE,R^2	• Time-series data for forecasting • Historical time-series data • Features such as trend and seasonality • External factors affecting time series	Not explicitly mentioned in the information
O. I. Oriekhoe, B. I. Ashiwaju, K. C. Ihemereze, U. Ikwue	Review of Big Data in FMCG Supply Chains	• Efficiency and Optimization Metrics • Cost-Benefit Analysis • Scalability and Adaptability	• Big data sources such as RFID, IoT, and sensors • Market-specific data for the African FMCG sector • Strategies employed by U.S. companies in FMCG supply chains	Not mentioned
S. Hwang, G. Yoon, E. Baek, B.-K. Jeon	Sales Forecasting Model for New-Released Products	Error Metrics: MAE, MSE, RMSE	Sales data for new-released and short-term mobile phones	95.85%

T. Huang, R. Fildes, D. Soopramanien	Competitive Information in FMCG Sales Forecasting	Error Metrics: MAE, MSE, RMSE	FMCG retail product sales data	90.34%
J. Henzel, M. Sikora	Gradient Boosting Application in FMCG Retail	Error Metrics: MAE, MSE, RMSE, Precision, Recall, F1-Score Area Under the Curve (AUC-ROC) Feature Importance Cross-Validation Techniques	• Performance indicators data for promotions • Historical promotion efficiency data • FMCG retail sales data • Product-specific attributes and features • External factors influencing promotion outcomes • Competitor information • Demographic data of target customer segments	Not explicitly mentioned
S. Gelper et al.	Identifying Demand Effects in Product Categories	Regression Analysis Cross-Category Demand Effects Causal Inference Methods Statistical Significance Tests Time-Series Analysis	Historical sales and purchase data for different product categories Pricing information, promotional strategies, and marketing efforts Consumer behavior data	90.25%
M. L. Demircan, E. Merdan	Order Prediction Methodology with Fuzzy Sets	Error Metrics: MAE, MSE, RMSE, Precision, Recall, F1-Score Area Under the Curve (AUC-ROC) Feature Importance Cross-Validation Techniques	Vendor-Managed Inventory System in FMCG Sector	95.56%
L. D. D. L. D. Dao, L. N. K. L. N. Khoi	Applying Deep Learning to Forecast Demand	Error Metrics: MAE, MSE, RMSE	Vietnamese FMCG Company	93.5%

3. METHODOLOGY

We are proposing the inclusion of Gradient Boost, another ensemble technique, to broaden the scope and potentially improve forecasting accuracy.

1. Examining XGBoost and Gradient Boost: Seeks to analyze the performance of XGBoost and Gradient Boost models, focusing on demand forecasting for FMCG data.

2. Methods for Extracting Features: Employs feature extraction techniques like Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) to pinpoint significant factors in predicting demand.

3. Comprehensive Testing and Assessment: Performs thorough testing of both models with FMCG data, assessing accuracy, reliability, and computational efficiency.

Understands the importance of demand forecasting in the FMCG industry to uphold ideal inventory levels and meet customer expectations. Concentrates on utilizing XGBoost and Gradient Boost, which are well-known as effective ensemble learning methods, to forecast demand in the FMCG industry. Introduces cutting-edge feature extraction techniques to boost model performance by identifying and leveraging the most crucial information in the dataset. Recognizes XGBoost as a widely recognized and powerful gradient-boosting method that is both rapid and scalable, serving as the standard approach.

The research focuses on the impact of feature extraction on FMCG demand forecasting, providing insights for decision-making, and improving supply chain management. It highlights the importance of understanding which model performs better in this sector, as it can enhance operational efficiency, reduce excess inventory, and improve resource allocation. The findings can also be generalizable, offering insights into the performance of XGBoost and Gradient Boost models in other forecasting domains.

4. RESULTS

Common metrics for binary classification are Accuracy, Precision, Recall, F1-score

Accuracy measures the overall correctness of the model.

Journal of Theoretical and Applied Information Technology
15th May 2024. Vol.102. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$ACC = \frac{TP + TN + FN}{TP + TN}$$

where: TP: True Positives, TN: True Negatives. FP: False Positives, FN: False Negatives

Precision (P) or Positive Predictive Value (PPV). Precision measures the accuracy of positive predictions.

$$P = \frac{TP}{TP + FP}$$

Recall (R) or Sensitivity or True Positive Rate (TPR). Recall measures the ability of the model to capture all positive instances.

MODEL	*RMSE for Rural Model*	*R²*	*RMSE for Urban Model*	*R²*
XgBoostML	1556.04524	0.982451302	1616.62442	0.977799808
GradiemtBoostML	197.0485439	0.999718585	71.01156065	0.999957165

$$R = \frac{TP}{TP + FN}$$

F1-score is the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{P \times R}{P + R}$$

ROC curve represents the trade-off between sensitivity and specificity. AUC-ROC measures the area under this curve. $AUC - ROC \in [0,1]$

For XGBoost and Gradient Boost, the overall objective function can be represented as:

$$\text{Objective Function} = \sum_{i=1}^{n} L(yi, \hat{yi}) + \sum_{k=1}^{k} \Omega(\text{fk})$$

Where $L(yi, \hat{yi})$ is is the logistic loss function for binary classification

$\Omega(\text{fk})$ is the regularization term

f_k represents the k-th tree in the ensemble.

Encode Categorical Features

Categorical features like `WH_capacity_size` and `zone` are encoded using Label Encoding to transform them into numeric format suitable for the machine learning model:

categorical_features = ['WH_capacity_size', 'zone']

After that we can Handle Missing Values.The code calculates medians for numeric columns only and fills missing values in these columns with their respective medians:

Mean= $\frac{Number\ of\ non\ missing\ values}{\sum \text{Non-missin values}}$

Median=Middle value of sorted non-missing values

Mode=Most frequently occurring non-missing value

Features that are thought to influence the target variable (`product_wg_ton`) are selected, and the datasets are split into training and testing sets. XGBoost regressor models and Gradient boost regression models for rural and urban data are initialized and trained.

S.no	*Model*	*Accuracy*	*Precision*	*Recall*	*AUC*
0	XGBoost	0.85	0.87	0.83	0.91
1	Gradient Boost	0.82	0.84	0.8	0.89

Table 1 : Comparison of performance metrics of XGBOOST and Gradient Boost models

Performance metrics for both Models

The models are evaluated using RMSE (Root Mean Squared Error) and R² (coefficient of determination) metrics to assess their accuracy and explanatory power

Table 2 : performance metrics of XGBOOST and Gradient Boost models for rural and urban

Scatter plots are generated to visually compare actual and predicted values for both rural and urban models.

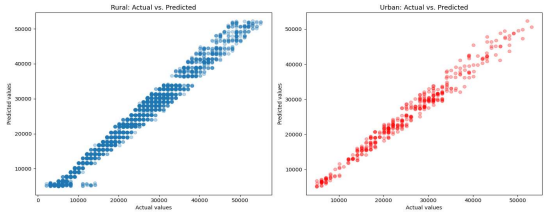

Figure:1 Visualization: Actual vs. Predicted Values (XgBoost ML)

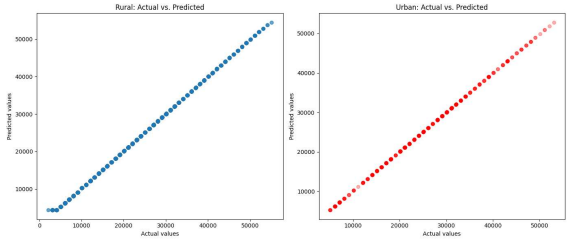

Figure:2 Visualization: Actual vs. Predicted Values (GBDT ML)

Histograms are plotted to analyze the distribution of prediction errors (differences between actual and predicted values)

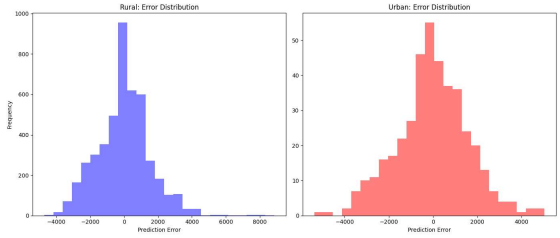

Figure 3 : Visualization: Error Distribution (XgBoost ML)

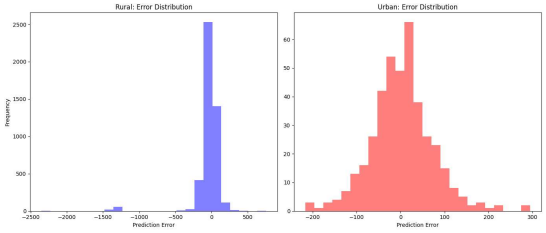

Figure 4 : Visualization: Error Distribution (GBDT ML)

This comprehensive approach not only builds and evaluates models for different subsets of the data (rural vs. urban) but also provides insights through visual analysis, helping to understand model performance and error characteristics in depth.

The RMSE values for the Gradient Boosting and XGBoost models, obtained through the ensemble technique, were found to be shown in the table.

MODEL	***RMSE for Rural Model***	***RMSE for Urban Model***
XgBoostML	1556.04524	1616.62442
GradiemtBoostML	197.0485439	71.01156065

Table 3 : RMSE Values for both Models

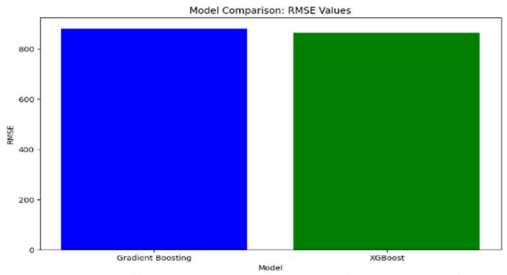

Figure 5: Comparison between the proposed and existing models

Lower RMSE values indicate better predictive accuracy, and in this case, the Gradient Boosting model demonstrated a lower RMSE compared to XGBoost, suggesting superior performance in terms of minimizing prediction errors.

5. CONCLUSION

This research proposes using Gradient Boost, an ensemble methodology, in addition to the known XGBoost method, to enhance demand forecasting in the fast-moving consumer goods (FMCG) industry. The study utilizes sophisticated feature extraction techniques, including principal component analysis (PCA) and recursive feature elimination (RFE), to improve model performance. We conducted a thorough assessment of the XGBoost and Gradient Boost models using FMCG data to evaluate their accuracy, dependability, and efficiency. The study's findings illuminate the efficacy of XGBoost and Gradient Boost in forecasting demand, offering significant insights for supply chain experts in the FMCG sector. By using gradient boost and feature extraction techniques, the goal is to reveal concealed patterns in the data, providing a deeper insight into the key elements influencing demand forecasting. This intelligence is essential for making well-informed choices and enhancing operational efficiency in the FMCG supply chain.

The research adds to the current literature by providing a comparative analysis of two influential ensemble learning algorithms in the context of demand forecasting for FMCG data. The study indicates that by combining XGBoost with Gradient Boost and feature extraction approaches, there is potential to expand the range and enhance the accuracy of forecasting.

The study's results are a significant resource for academics and practitioners, helping them make educated decisions on the most appropriate methodologies for their demand forecasting requirements. Supply chain experts in the FMCG sector may improve their forecasting capacities and decision-making by using the insights obtained from this research. This is anticipated to enhance operational efficiency and decrease costs in the FMCG supply chain.

6. FUTURE ENHANCEMENT

In the future, this research lays the foundation for several promising avenues in the realm of demand forecasting for the fast-moving consumer goods (FMCG) industry. Firstly, the integration of cutting-edge deep learning techniques, such as recurrent neural networks (RNNs) or attention mechanisms, could be explored to further refine the model's capacity to capture intricate temporal dependencies. Additionally, the dynamic nature of market conditions could be addressed by developing adaptive feature selection strategies or even implementing real-time forecasting capabilities. The continued pursuit of enhanced model explain ability remains crucial, with the potential exploration of advanced techniques in explainable AI to ensure transparency and foster trust among stakeholders. Furthermore, assessing the cross-industry applicability of the ensemble learning approach and conducting continuous evaluations and updates to the model will contribute to its longevity and relevance. Collaboration with industry experts, the exploration of additional external data streams, and the development of user-friendly interfaces can collectively propel this research towards practical

implementation, offering sustainable benefits for supply chain experts in the FMCG sector. Ultimately, the future scope involves a comprehensive and evolving approach that integrates emerging technologies, addresses real-world challenges, and aligns with the evolving landscape of demand forecasting in the FMCG industry.

REFERENCES:

[1] E. Tarallo *et al.*, "Machine Learning in Predicting Demand for Fast-Moving Consumer Goods: An Exploratory Research," *IFAC-PapersOnLine*, 2019, doi: 10.1016/j.ifacol.2019.11.203.

[2] P. M. Pardalos, R. J. Hyndman, R. J. Hyndman, R. J. Hyndman, Y. Khandakar, and Y. Khandakar, "Automatic Time Series Forecasting: The forecast Package for R," *Journal of Statistical Software*, 2008, doi: 10.18637/jss.v027.i03.

[3] O. I. Oriekhoe, B. I. Ashiwaju, K. C. Ihemereze, and U. Ikwue, "REVIEW OF BIG DATA IN FMCG SUPPLY CHAINS: U.S. COMPANY STRATEGIES AND APPLICATIONS FOR THE AFRICAN MARKET," International Journal of Management & Entrepreneurship Research, 2024, doi: 10.51594/ijmer.v6i1.711.

[4] A. Mebal.P* et al., "Predicting the Demand for Fmcg using Machine Learning," International Journal of Engineering, 2021, doi: 10.35940/ijeat.c2253.0210321.

[5] P. McCullagh, P. McCullagh, J. A. Nelder, and J. A. Nelder, "Generalized Linear Models," null, 1983, doi: null.

[6] S. Makridakis and S. Makridakis, "The art and science of forecasting An assessment and future directions," International Journal of Forecasting, 1986, doi: 10.1016/0169-2070(86)90028-2.

[7] A. Krishna et al., "Sales-forecasting of Retail Stores using Machine Learning Techniques," null, 2018, doi: 10.1109/csitss.2018.8768765.

[8] S. Hwang, G. Yoon, E. Baek, and B.-K. Jeon, "A Sales Forecasting Model for New-Released and Short-Term Product: A Case Study of Mobile Phones," null, 2023, doi: 10.3390/electronics12153256.

[9] T. Huang, T. Huang, R. Fildes, R. Fildes, D. Soopramanien, and D. Soopramanien, "The value of competitive information in forecasting FMCG retail product sales and the variable selection problem," European Journal of Operational Research, 2014, doi: 10.1016/j.ejor.2014.02.022.

[10] J. Henzel, J. Henzel, M. Sikora, M. Sikora, and M. Sikora, "Gradient Boosting Application in Forecasting of Performance Indicators Values for Measuring the Efficiency of Promotions in FMCG Retail," Conference on Computer Science and Information Systems, 2020, doi: 10.15439/2020f118.

[11] S. Gelper et al., "Identifying demand effects in a large network of product categories," Journal of Retailing, 2016, doi: 10.1016/j.jretai.2015.05.005.

[12] J. H. Friedman, J. H. Friedman, and J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," Annals of Statistics, 2001, doi: 10.1214/aos/1013203451.

[13] R. Fildes et al., "Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning," International Journal of Forecasting, 2009, doi: 10.1016/j.ijforecast.2008.11.010.

[14] M. L. Demircan, M. L. Demircan, M. L. Demircan, M. L. Demircan, E. Merdan, and E. Merdan, "A Proposed Order Prediction Methodology for Vendor-Managed Inventory System in FMCG Sector Based on Interval-Valued Intuitionistic Fuzzy Sets," International Journal of Computational Intelligence Systems, 2021, doi: 10.2991/ijcis.d.210423.004.

[15] L. D. D. L. D. Dao and L. N. K. L. N. Khoi, "Applying deep learning to forecast the demand of a Vietnamese FMCG company," Tạp Chí Khoa Học Trường Đại Học Quốc Tế Hồng Bàng, 2023, doi: 10.59294/hiujs.vol.5.2023.552.

[16] L. G. Cooper et al., "Promocast: a New Forecasting Method for Promotion Planning," Marketing Science, 1999, doi: 10.1287/mksc.18.3.301.

[17] C. W. Chu et al., "A comparative study of linear and nonlinear models for aggregate retail sales forecasting," International Journal of Production Economics, 2003, doi:10.1016/s0925-5273(03)00068-9.

[18] V. singh chandraul, V. singh chandraul, S. K. Barode, and S. kumar barode, "Optimizing Demand and Forecasting in Supply Chain Management Using AI," SMART MOVES JOURNAL IJOSCIENCE, 2018, doi: 10.24113/ijoscience.v4i12.174.

[19] M. Blachnik, M. Blachnik, J. Henzel, and J. Henzel, "Estimating the performance indicators of promotion efficiency in fmcg retail," null, 2020, doi: 10.1007/978-3-030-63833-7_27.

[20] Z. Y. Alzubaidi and Zya, "A Comparative Study on Statistical and Machine Learning Forecasting Methods for an FMCG Company," null, 2021, doi: null.