# Data Analytics for Product Segmentation and Demand Forecasting of a Local Retail Store Using Python

Arun Kumar Mishra[1], Megha Sinha[2]
Department of Computer Science and Engineering, University College of Engineering and Technology (UCET), Vinoba Bhave University, Hazaribag -825301, Jharkhand, India[1]
Department of Computer Science and Engineering, Sarala Birla University, Ranchi-835103, Jharkhand, India[2]

***Abstract*—In today's competitive business environment, understanding customers' expectations and choices is a necessity for the successful operations of a retail store. Forecasting demand also plays an important role in maintaining inventory at an optimum level. The work utilises data analytics for product segmentation and demand forecasting in a local retail store. Python is being used as a programming language for data analytics. Historical sales data of a local store has been used to categorise products into different segments. Statistical techniques and a k-means clustering algorithm have been used to understand different segments of the product. Machine learning algorithms and time series models have been used to forecast future sales trends. The business insights allow the retail store to meet customers' expectations, manage inventory at an optimum level and enhance supply chain efficiency. The present work seeks to illustrate how data-driven tactics can enhance operational decision-making in retail.**

***Keywords*—Data analytics; product segmentation; demand forecasting; multicriteria ABC classification; seasonality**

## I. Introduction

To ensure business growth and maintain operational efficiency, understanding customer choices and forecasting demand for various products have become important in the current competitive retail environment. The problems a local retail store faces include but are not limited to inventory management, optimisation of sales strategy and fulfilling customers' expectations in changing market trends. In this scenario, product segmentation and demand forecasting help overcome these hurdles to run a successful business.

Based on common characteristics such as sales performance, revenue generation, demand trends and consumer preferences, products are classified into different groups. This process is nothing but product segmentation. It helps retailers customise marketing strategies, optimise inventory, and enhance overall resource allocation. Previous sales data is utilised to predict future demand trends in demand forecasting. It helps retailers make proactive decisions in procurement, inventory management, and supply chain operations.

Python, an object-oriented programming language, provides a rich set of libraries and tools for data analytics. Python provides extensive solutions for addressing intricate retail difficulties, encompassing data pre-treatment, visualisation, machine learning, and time series modelling. Pandas, NumPy, Scikit-learn, and Prophet are particularly adept at product clustering, trend analysis, and prediction modelling.

This article examines the utilisation of Python-based data analytics methods for efficient product segmentation and demand forecasting in a small retail establishment. The study seeks to analyse previous sales data to Determine specific product segments for focused marketing and inventory approaches and construct predictive models to anticipate future demand, reducing stockouts and excess inventory.

This study's findings emphasise that data-driven techniques can enhance decision-making processes in retail, resulting in greater efficiency, customer satisfaction, and profitability.

## II. Literature Review

Generally, uniform control measures for all inventory products are inadvisable. The high-value items may be essential to the viability of the firm. The study in [1] talked about ABC and multicriteria ABC analysis. In ABC analysis, products are classified into three classes: A, B and C. Class A items entail significant stock-out expenses and necessitate stringent control measures. It emphasised that multicriteria ABC analysis was crucial for comprehending different product categories: volume drivers, margin drivers, regular movers, and slow movers. In multicriteria analysis, categories A_B and A_C denote volume driver items, B_A and C_A indicate margin driver items, categories B_B, B_C, and C_B imply regular items, category C_C reveals slow-moving items, and A_A encompasses both margin and volume driver items. The research performed multicriteria ABC analysis on the online retail dataset utilising data analytics methodologies. The study in [2] proposed using a three-phased Multi-Criteria Inventory Classification (MCIC) integrating the Analytical Hierarchy Process (AHP), Fuzzy C-Means (FCM) algorithm, and a newly proposed Revised-Veto (RVeto) phase to adhere to the ABC Classification principles and enhance its application and adaptability. Classification based on several criteria is essential to meeting management's needs in the current context. The study in [3] presented a semi-supervised explainable methodology that integrated semi-supervised clustering with explainable artificial intelligence. The semi-supervised method integrated intelligent initialisation with a constrained clustering process that directed the classification procedure towards Pareto-distributed items. At the

same time, explainable artificial intelligence was employed to generate comprehensive micro and macro explanations of inventory categories at both the item and class levels. Implementing the suggested method for the automatic classification of chemical items within a distribution organisation has demonstrated its efficacy in delivering precise, transparent, and thoroughly elucidated ABC classifications. The study in [4] presented an optimal multi-criteria ABC inventory classification for supermarkets to manage commodities based on unit price, lead time, and annual usage. Of the 442 objects, 30 were categorised as group "A," 31 as group "B," and 27 as group "C" under the new ABC classification; nevertheless, all these things were categorised in group "A" in the conventional ABC classification. The study in [5] indicated that AI-based methods exhibited more accuracy than multiple discriminant analysis (MDA). The statistical study specified that SVM facilitated superior classification accuracy compared to alternative AI methodologies. This discovery indicated the potential for employing AI-driven methodologies for multi-criteria ABC analysis within enterprise resource planning (ERP) systems. The study in [6] aimed to present a case-based multiple-criteria ABC analysis that enhances the traditional method by incorporating other factors, such as lead time and SKU criticality. It offered greater managerial flexibility. Decisions from instances served as input, with preferences for alternatives represented naturally using weighted Euclidean distances. It facilitated easy understanding for the decision-maker. The study in [7] examined current portfolio models in procurement that categorise purchases into several product classifications. Case studies from two European automotive OEMs and two vehicle industry suppliers and benchmarking interviews at Toyota, Japan, were used to establish a connection between these product categories and various supplier types. Further, it tried correlating the product categories and supplier types with the specification process—specifically, associating the specification types with their respective generators. The study in [8] conducted supplier segmentation within the automobile sector and proposed four techniques for supplier relationships. Additionally, a four-phase approach for analysing, selecting, and managing decisions on a dynamic relationship strategy with suppliers in the automotive sector was delineated. The study in [9] reviewed the available literature, focusing on market conditions, supplier characteristics, buyer characteristics, and the connections between buyers and suppliers. The study in [10] formulated an innovative methodology for supplier segmentation. Fuzzy logic was utilised to divide suppliers in a broiler firm.

The study in [11] performed a comparative analysis of machine learning algorithms for demand forecasting under uncertainty. The research utilised a synthetic dataset. The machine learning algorithms compared were Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Machine Regression (SVR), XG Boost Regression on the parameters of Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The study in [12] proposed a model that integrated time series analysis, boosting and deep learning for demand forecasting. It achieved a significant enhancement in accuracy relative to state-of-the-art studies. The testing utilised authentic data from Turkey's SOK Market. The article compared the Decision Tree Classifier, Gaussian Naive Bayes, and K-Nearest Neighbours (KNN). The Gaussian Naive Bayes technique exhibited the greatest accuracy in demand estimation. The study in [13] focused on demand forecasting and consumer satisfaction within the retail sector. It emphasised the importance of precise demand estimation for merchants. The discussion encompassed machine learning methodologies for forecasting product demand. The paper considered variables for prediction, including time, location, and historical data.

## III. Present Work

In this paper, product segmentation was performed using ABC and Multicriteria ABC analysis. First, data was collected, and then it was prepared for the segmentation exercise. Then, segmentation was performed using Python's inventorize package. After that, classification algorithms were applied to it, and performance was evaluated. The flow of work is shown in Fig. 1.

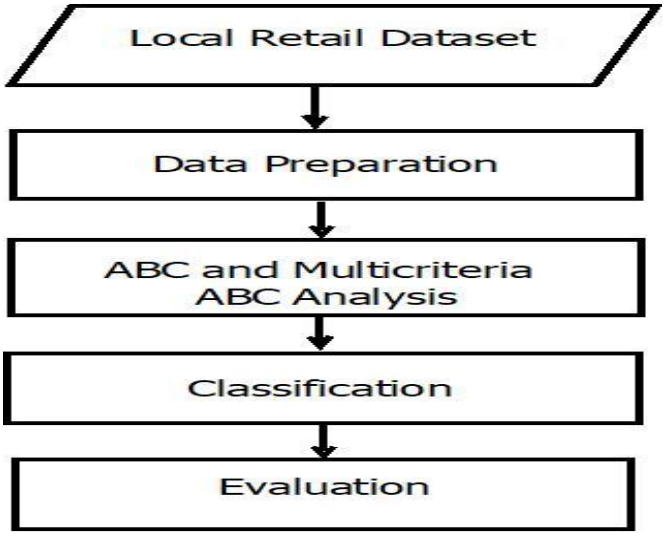



Fig. 1. Workflow for product segmentation.

The same dataset was analysed for demand trends, and a comparative analysis of different machine learning algorithms was done to forecast the demand for A-class products. Fig. 2 shows the workflow for this.

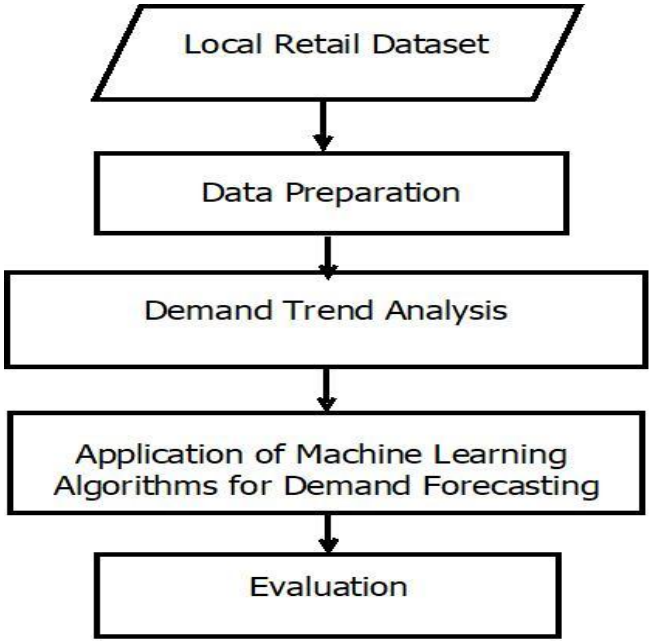



Fig. 2. Workflow for demand trend analysis and forecasting.

Two years monthly sales data of a local retail store situated at Hazaribag was captured and stored in a file named 'Sales_data.xlsx'. The dataset sample is displayed in Fig. 3.

```
[ ] data=pd.read_excel('Sales_data.xlsx')
```

```
[ ] data.head()
```

| | Month | SKU | Quantity | Price | Revenue |
|---|---|---|---|---|---|
| 0 | 2022-04 | Mask | 12.0 | 50.00 | 600.00 |
| 1 | 2022-04 | N95 Mask | 12.0 | 50.00 | 600.00 |
| 2 | 2022-04 | Refill | 5.0 | 508.54 | 2542.70 |
| 3 | 2022-04 | Refilling ABC & Servicing | 5.0 | 508.54 | 2542.70 |
| 4 | 2022-04 | SHOES | 12.0 | 452.38 | 5428.56 |

Fig. 3. Local retail dataset sample.

The dataset has five columns: 'Month', 'SKU', 'Quantity', 'Price', and 'Revenue'. It has 4339 records and was analysed for null values and duplicates. After removing records with null values and eliminating duplicates, the dataset comprised 4089 rows. It was further analysed for quantity value. Only those records were kept in which the quantity value was greater than 0. For analysis purposes, one new column, 'Date', was added using the column 'Month', converting it to a datetime object and dropping it for further analysis. Fig. 4. shows the sample prepared dataset. The pertinent columns of the dataset, namely 'SKU', 'Quantity', and 'Revenue', were retained for ABC and multicriteria ABC analysis. Additionally, data was consolidated according to 'SKU'.

```
[ ] data.set_index('Date', inplace=True)
```

```
[ ] data.head()
```

| Date | SKU | Quantity | Price | Revenue |
|---|---|---|---|---|
| 2022-04-01 | Mask | 12.0 | 50.00 | 600.00 |
| 2022-04-01 | N95 Mask | 12.0 | 50.00 | 600.00 |
| 2022-04-01 | Refill | 5.0 | 508.54 | 2542.70 |
| 2022-04-01 | Refilling ABC & Servicing | 5.0 | 508.54 | 2542.70 |
| 2022-04-01 | SHOES | 12.0 | 452.38 | 5428.56 |

Fig. 4. Sample local retail dataset.

The inventorize module in Python was utilised to conduct an ABC analysis based on volume. Further, the ABC analysis was performed on revenue. After that, multicriteria analysis based on 'Revenue' and 'Quantity' was performed on the dataset. The dataset obtained after this analysis was used to perform a comparative study of machine learning algorithms viz. KNN, Decision Tree, Random Forest and Naïve Bayes algorithm for classification. The product mix categorisation was treated as the labelled output, while the rest of the columns were used as the basis for classification. First of all, data was split into training and test data. The test size of the data was kept at 20% of the total data. The models were trained and then tested. Confusion matrices were plotted for each model, and accuracy scores were calculated.

As shown in Fig. 4, the prepared dataset was used to understand the demand trend. A graph was plotted for monthly sales over time to know monthly sales trends. Seasonal decomposition was performed to learn more about seasonality trends, and a graph was plotted. Further analysis was carried out to compare machine learning algorithms, viz. Linear regression, Decision tree, Random forest, SVR and XG Boost regressor for demand forecasting of class A items of local retail store. Next, demand forecasting was performed using the said machine learning algorithms. The dataset was split into train and test data with 20% data size. A comparison of predictions was done for all these items. The performance metrics included MAE, MSE and RMSE. To visualise the results in a single frame, grouped bar graphs were plotted for MAE, MSE, and RMSE. Then, the graph was plotted to visualise the comparative performance of the machine learning algorithm in this study.

## IV. Results and Discussions

The result counts of ABC analysis on volume and revenue are displayed in Fig. 5 and Fig. 6, respectively.

Then, multicriteria analysis based on 'Revenue' and 'Quantity' was performed on the dataset. Fig. 7 displays the product mix count.

The dataset obtained after this analysis was used to perform a comparative study of machine learning algorithms, viz., KNN, Decision Tree, Random Forest, and Naïve Bayes algorithm for classification. Fig. 8 shows the confusion matrices for these algorithms.

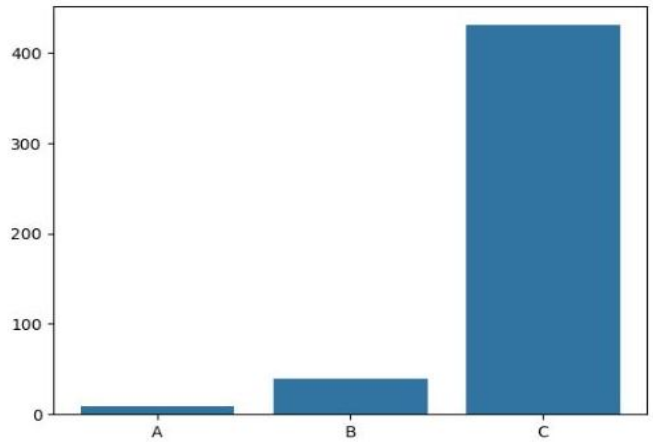



Fig. 5. ABC Count by volume.

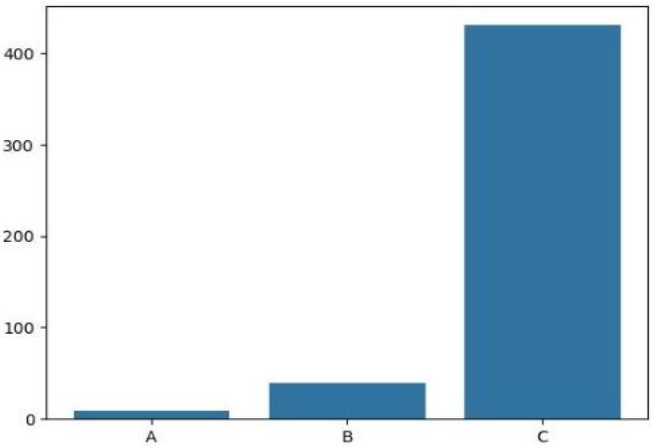



Fig. 6. ABC Count by revenue.

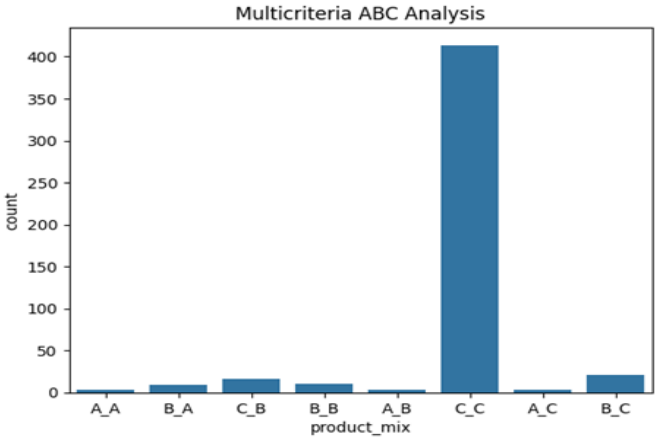



Fig. 7. Multicriteria ABC analysis.

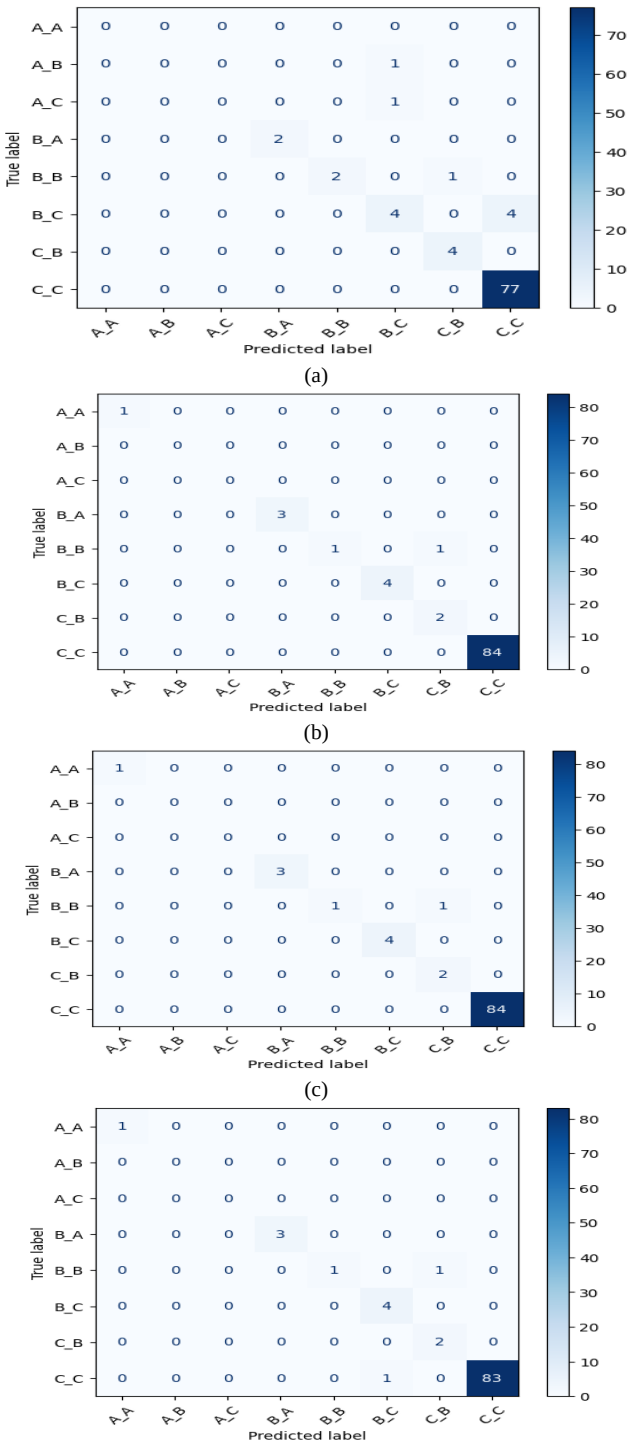



Fig. 8. Confusion matrix (a) KNN (b) Decision Tree (c) Random Forest and (d) Naïve Bayes classification algorithms.

The comparative chart for the accuracy scores has been displayed in Fig. 9.

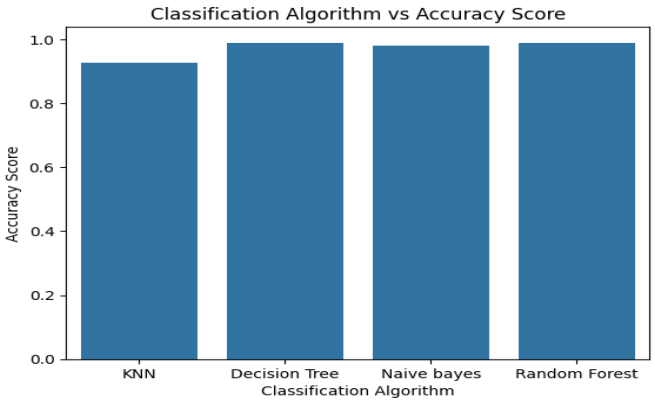



Fig. 9. Accuracy scores comparison for classification.

Decision Tree and Random Forest classification algorithms with identical accuracy scores outperformed KNN and Naïve Bayes algorithms.

Next, a graph plotting monthly sales over time shows monthly sales trends. Fig. 10 displays the seasonal decomposition for the same.

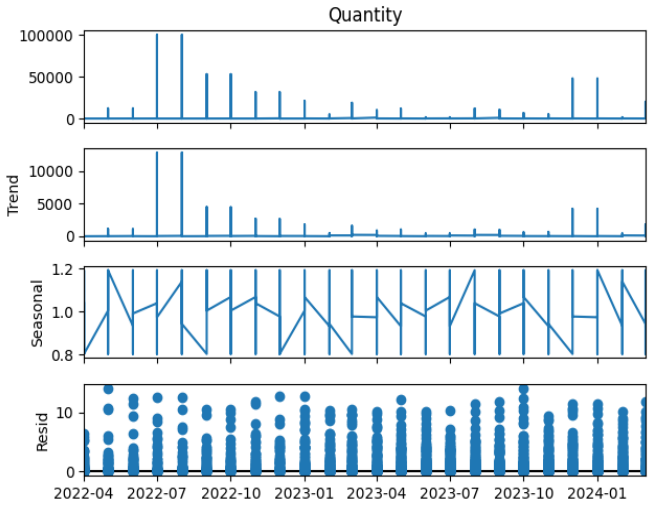
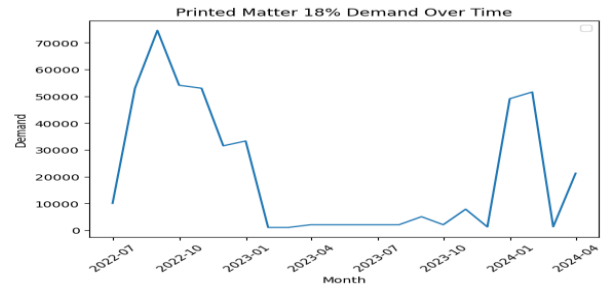
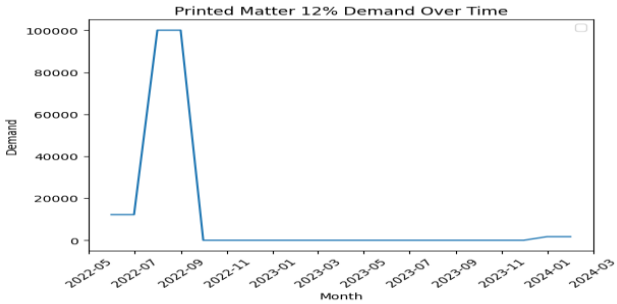



Fig. 10. Seasonality trend for monthly sales data.

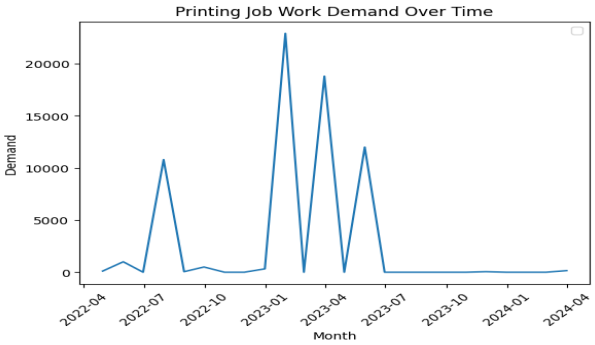
There are irregular spikes in this plot, suggesting an occasional high level of demand at irregular periods. It is clear from the trend plot that there is variability in the dataset. It indicates that the variability is inconsistent enough to make an upward or downward trend over time. The seasonal component of the graph indicates a continuous seasonal influence. However, that influence is very thin. It is shown by minor variations in numbers within a recurring period. The residuals in the plot are scattered, showing greater fluctuations during instances of spikes in the data. It suggests that the spikes may not be described by the trend or seasonality. Further analysis was carried out to compare machine learning algorithms, viz. Linear regression, Decision tree, Random Forest, SVR and XG Boost regressor for demand forecasting of class A items of local retail store. Fig. 11 shows the demand pattern for these items.



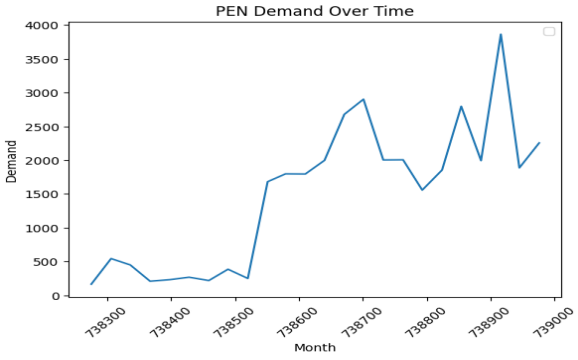

(a)



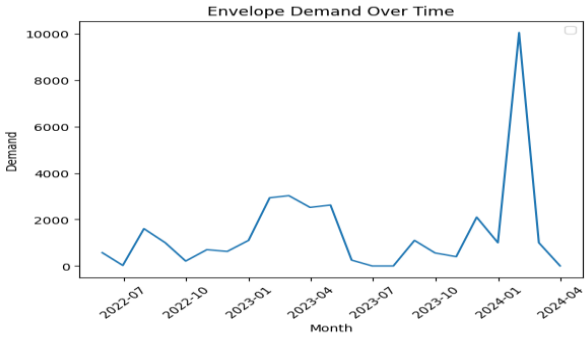

(b)



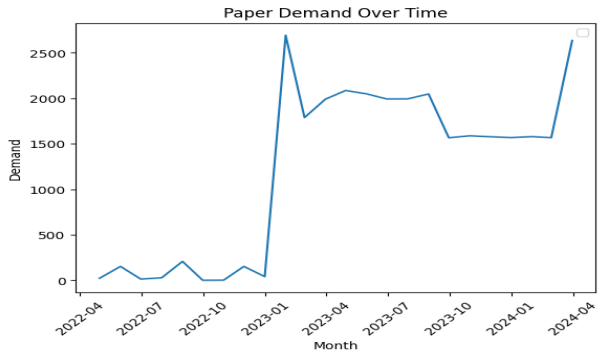

(c)



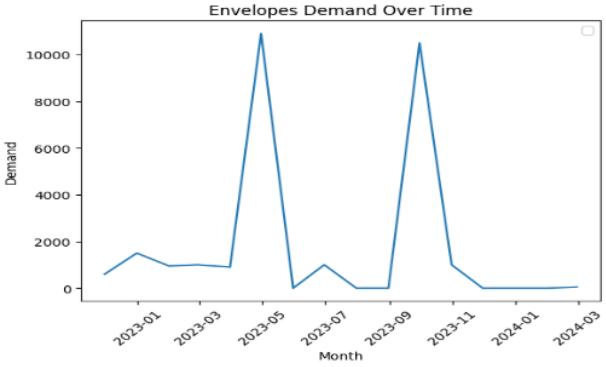

(d)



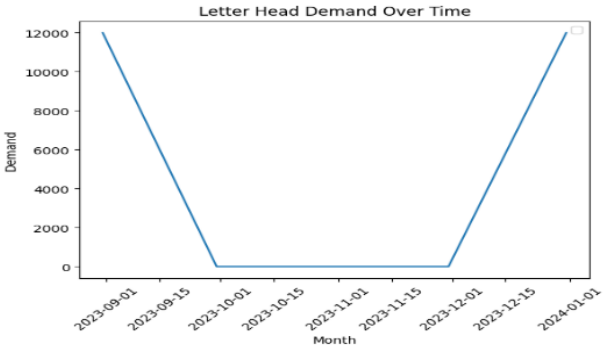

(e)



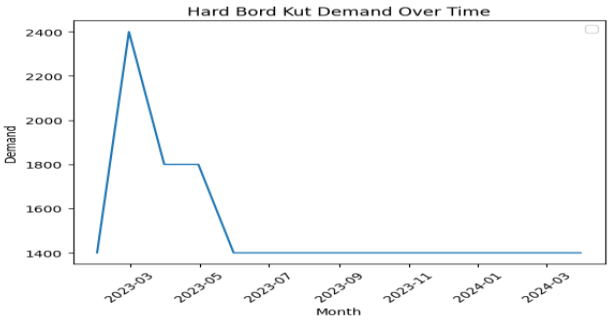

(f)




(g)




(h)




(i)

Fig. 11. Demand pattern for (a) Printed Matter 18% (b) Printed Matter 12% (c) Printing Job Work (d) PEN (e) Envelope (f) Paper (g) Envelopes (h) Letter Head (i) Hard Bord Kut.

Demand for these products was forecasted using machine learning algorithms, and MAE, MSE, and RMSE were calculated. To visualise the results in a single frame, grouped bar graphs have been plotted for MAE, MSE, and RMSE, as shown in Fig. 12, 13, and 14, respectively.

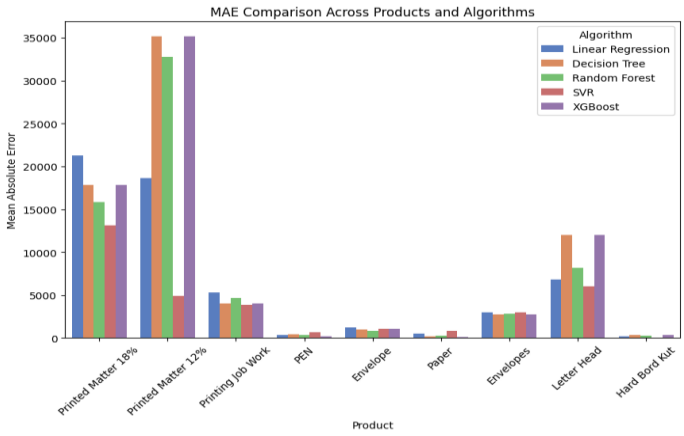



Fig. 12. Comparative performance of machine learning algorithms based on MAE for all class 'A' items.

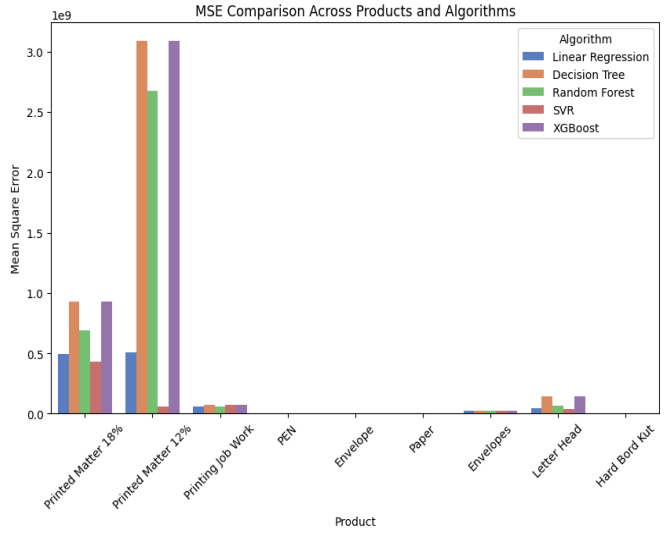



Fig. 13. Comparative performance of machine learning algorithms based on MSE for all class 'A' items.

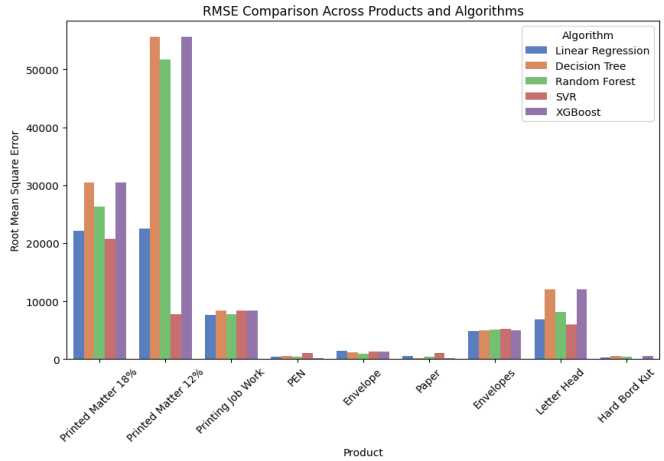



Fig. 14. Comparative performance of machine learning algorithms based on RMSE for all class 'A' items.

Fig. 15 shows the comparative performance of the machine learning algorithm compared in this study.

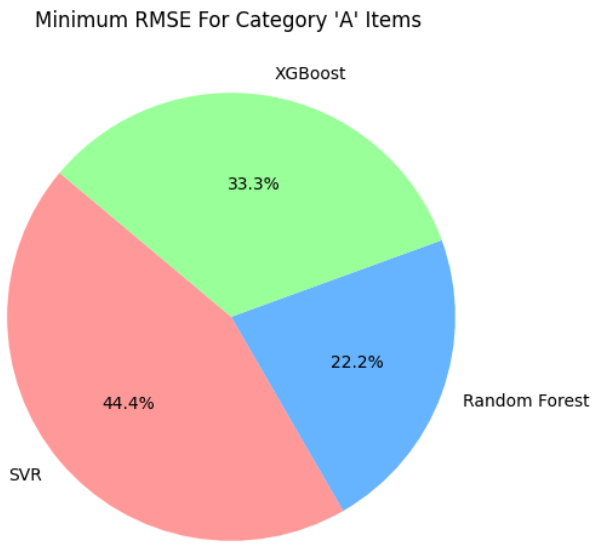



Fig. 15. Comparative analysis of best-performing machine learning algorithms for demand forecasting.

It can be seen that SVR outperformed other algorithms for 44.4% of class A items, XGBoost performed better than other algorithms for 33.3% of items, and Random Forest outperformed other algorithms for 22.2% of class A items.

## V. Conclusion

Given that clients seek a diverse range of products while desiring more excellent value for their expenditure, it is imperative to comprehend the several categories of products, including volume drivers, margin drivers, and frequent and slow movers. Multi-criteria ABC analysis serves as an effective tool for conducting this segmentation. This study aims to evaluate the efficacy of classification algorithms in conducting multi-criteria ABC analysis on a retail dataset. Products have been classified based on 'Quantity' and 'Revenue' parameters into A_A, B_A, C_B, B_B, A_B, C_C, A_C and B_C categories. In the contemporary internet company landscape, categorising products and locating providers of vital commodities has become crucial for business viability. Consequently, strategic collaborations may be established for commodities that generate volume and margin. The 'inventorize' module was an effective Python tool for conducting multi-criteria analysis based on quantity and income. It may assist in determining the critical elements to retain. The results indicate that, among the classification algorithms evaluated based on accuracy score, Random Forest and Decision Tree Classifier exhibited comparable performance. Further, a data-driven approach was applied to identify demand trend analysis and demand forecasting on class A items of a local retail store. Machine learning algorithms were also compared to forecast these items. SVR outperformed other algorithms for nearly half of the products in this respect.

## References

[1] A. K. Mishra, & M. Sinha. Data Analytics for Multi-criteria ABC Analysis and Supplier Segmentation in Making a Competitive Supply Chain Using Python. In PROCEEDINGS OF 10th INTERNATIONAL SYMPOSIUM ON FUSION OF SCIENCE AND TECHNOLOGY (ISFT-2024) JANUARY 4-8, 2024. https://www.jcboseust.ac.in/assets/files/sovenir_PROCEEDING_ISFT_2024_1.pdf

[2] Fatih Yiğit, Sakir Esnaf, "A New Fuzzy C-Means and AHP-Based Three-Phased Approach for Multiple Criteria ABC Inventory Classification." IMSS'19 Sakarya University - Sakarya/Turkey, 9-11 September 2019, pp. 633-642

[3] A. A. Qaffas, M. A. B. Hajkacem, C. -E. B. Ncir and O. Nasraoui, "Interpretable Multi-Criteria ABC Analysis Based on Semi-Supervised Clustering and Explainable Artificial Intelligence," in IEEE Access, vol. 11, pp. 43778-43792, 2023, doi: 10.1109/ACCESS.2023.3272403.

[4] Aregawi Yemane, Alehegn Melesse Semegn and Ephrem Gidey," ABC Classification for Inventory Optimization (Case Study Family Supermarket)", Industrial Engineering & Management, Research - (2021) Volume 10, Issue 5, ISSN: 2169-0316

[5] Min-Chun Yu (2011). Multi-criteria ABC analysis using artificial-intelligence-based classification techniques. Expert Syst. Appl.. 38. 3416-3421. 10.1016/j.eswa.2010.08.127.

[6] Ye Chen, Kevin W. Li, D. Marc Kilgour, Keith W. Hipel, A case-based distance model for multiple criteria ABC analysis, Computers & Operations Research, Volume 35, Issue 3, 2008, Pages 776-796, ISSN 0305-0548,https://doi.org/10.1016/j.cor.2006.03.024.

[7] R. Nellore, and K. Söderquist (2000) 'Portfolio approaches to procurement: Analysing the missing link to specifications', Long Range Planning, 33, 245-267.

[8] G. Svensson (2004) 'Supplier segmentation in the automotive industry: A dyadic approach of a managerial model', International Journal of Physical Distribution and Logistics Management, 34, 12-38.

[9] M. Day, G. M. Magnan and M. M. Moeller (2010) 'Evaluating the bases of supplier segmentation: A review and taxonomy', Industrial Marketing Management, 39, 625-639.

[10] J. Rezaei and R. Ortt, (2013) 'Multi-criteria supplier segmentation using a fuzzy preference relations based AHP', European Journal of Operational Research, 225, 75-84.

[11] Arun Kumar Mishra, Megha Sinha, & Sudhanshu Kumar Jha. (2024). Comparative analysis of machine learning algorithms for demand forecasting under uncertainty. Computer Science & IT Research Journal, 5(8), 1817-1827. https://doi.org/10.51594/csitrj.v5i8.1409.

[12] Z. H. Kilimci, A. O. Akyuz, M. Uysal, S. Akyokus, M. O. Uysal, B. Atak Bulbul & M. A. Ekmis (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. Complexity, 2019(1), 9067367.

[13] A. I. Arif, S. I. Sany, F. I. Nahin, & A. S. A. Rabby (2019, November). Comparison study: product demand forecasting with machine learning for shop. In 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 171-176). IEEE.