International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 1, Jan-Feb 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Sales Analytics and Optimization: A Comprehensive Study Using Transactional Data

Dr. Gaytri Devi[1], Dr. Komal Sehrawat [2,] Shivangi Dixit[3]
GVM Institute of Technology & Management, DCRUST University, Murthal
[1]gayatri.dhingra1@gmail.com, [2]komalsehrawatbpsmv@gmail.com, [3]adv.shivangidixit@gmail.com

----------**********----------

Abstract:

Sales analytics is the process of collecting, analyzing, and interpreting sales data to gain insights into sales performance, customer behavior, market trends, and business opportunities. It helps businesses make data-driven decisions to improve sales strategies, optimize resources, forecast sales, and enhance overall profitability.With the exponential growth of e-commerce, accurate sales forecasting and trends have become strategic imperatives to stay competitive in a volatile market. The study in this paper was undertaken to explore how data-driven techniques can optimize retail sales operations and customer engagement. We utilized a comprehensive dataset of historical transactions from an online retail platform and applied statistical and machine learning models to analyze sales trends, product performance, geographic distributions, and customer behaviors. The study employed various forecasting models, including a time series model ARIMA and Machine learning models such as Linear Regression and K-Nearest Neighbors comparing their performance in terms of accuracy. It has been found that ML model performs better than Time series model. Additionally, RFM (Recency, Frequency, Monetary) analysis and K-Means clustering were used for customer segmentation. The findings aim to guide retailers on where to focus their marketing, inventory, and customer retention strategies for improved profitability.

Keywords **—Sales Forecasting, RFM, KNN, Linear Regression, ARIMA, K-means, Data Analytics**

----------**********----------

I. INTRODUCTION

The retail industry is undergoing rapid digital transformation, with e-commerce playing a dominant role in consumer purchasing behavior. One of the most significant challenges for online retailers is managing sales volatility due to seasonality, market trends, and customer preferences Effective sales forecasting helps businesses anticipate demand, avoid stockouts or overstocking, and allocate resources more efficiently. When combined with customer segmentation, forecasting allows businesses to personalize offerings, improve customer satisfaction, and increase lifetime value.

This research focuses on developing and comparing forecasting models, coupled with sales trends and behavioral customer segmentation, to derive actionable insights for online retail optimization.

The structure of the paper is organized as follows: Section 2 presents the literature review, Section 3 outlines the methodology, Section 4 focuses on data analytics and visualization, Section 5 discusses the business insights derived from the analysis, and Section 6 concludes the paper with key findings and future directions.

2. Literature Review

Sales analytics and forecasting have become crucial for strategic decision-making in the retail sector, especially with the rapid rise of e-

commerce.Traditionally, statistical models such as the Autoregressive Integrated Moving Average (ARIMA) have been widely used to model and forecast sales data due to their strength in capturing linear temporal dependencies. For example, ARIMA has been shown to effectively handle seasonal and trend components in retail time series (Ensafi et al., 2022). Similarly, Sharma, Patel, & Gupta (2021) highlighted how ARIMA models provide strong baselines for short-term sales forecasting, especially when historical data is stationary and exhibits consistent patterns.

However, with the growing complexity of consumer behavior and promotional cycles, machine learning models have gained traction for their ability to model non-linear relationships and interactions. Algorithms such as linear regression, K-nearest neighbors (KNN), and tree-based methods outperform classical approaches in many retail contexts. For example, Pavlyshenko (2019) demonstrated that machine learning methods, including boosted decision trees and ridge regression, improve forecasting accuracy by incorporating promotional and external features. Customer segmentation has gained prominence as a complementary analytic strategy to forecasting. The Recency, Frequency, and Monetary (RFM) analysis provides a foundational approach to understand customer behavior by categorizing them based on their purchasing patterns. RFM-based segmentation allows businesses to identify high-value customers and tailor personalized marketing interventions (Anitha & Patil, 2022).Similarly, Hallimsyore Kalegowda (2023) demonstrated that combining RFM analysis with K-means clustering enhances the understanding of customer lifetime value and supports proactive inventory management. Data Driven Analytics presented by Kadam and Vhatkar (2022), who highlight the advantage of integrating ML techniques to accommodate the seasonal and promotional dynamics influencing sales.

Overall, the confluence of advanced machine learning techniques, robust segmentation frameworks like RFM and K-means clustering, and transactional data analytics provides a comprehensive foundation for sales optimization. The integrated approach not only improves forecasting accuracy but also facilitates more effective customer engagement and profitability strategies.

3. Methodology

This study is based on an online retail dataset comprising 541,909 transactions sourced from Kaggle. The data spans a one-year period from 2010 to 2011, and includes key fields such as: InvoiceDate (Transaction date) ,Description (Productname) ,Quantity ,UnitPrice ,CustomerID Country. For the findings to be accurate and reliable, it is essential that the dataset is thoroughly cleaned and preprocessed[4].The raw dataset was preprocessed to ensure reliability and accuracy for analysis. Data was Cleaned to remove returns, null values, and outlier.

Key Analytical Techniques

A) **Time Series Analysis** : Time series techniques were applied to uncover trends and seasonality in sales over time, across geographical regions, and product categories. This analysis helped in identifying patterns useful for forecasting and business planning.

B) ***Correlation Analysis:*** To understand relationships between numerical variables, a correlation matrix was constructed and visualized using heatmaps. This approach was chosen for its interpretability, especially for business stakeholders, as opposed to more abstract techniques like PCA.

C) **Customer Segmentation – RFM and K-Means Clustering :**To segment customers, RFM (Recency, Frequency, Monetary) analysis was conducted to quantify customer behavior. Based on the RFM scores, K-Means clustering was applied to categorize customers into: High Spenders ,Medium Spenders ,Low and Occasional Buyers. K-Means was selected for its simplicity, scalability, and superior efficiency in handling large datasets, compared to hierarchical clustering.

D) Predictive Modeling – Sales Forecasting: To forecast future sales, three models were implemented and compared:ARIMA – a statistical time series model, Linear Regression – a fundamental machine learning regression model and K-Nearest Neighbors (KNN) Regressor – a distance-based, non-parametric model. The performance of the forecasting models was evaluated using the metrics:Absolute Error (MAE),Root Mean Squared Error (RMSE),Coefficient of Determination (R^2 Score).These metrics provided insights into model accuracy and predictive capabilities, enabling effective comparison across approaches.

4. Data Analysis and Visualization

In this section, we present the visualizations and conclusions derived from the comprehensive analysis performed, including sales trends, customer behavior, product performance as well as predictive modeling for sales forecasting.

4.1 Sales Trend and Seasonality

The analysis reveals sales fluctuations during the first half of the year, with noticeable dips in February and April. Starting in September, there is a sharp increase in sales, reaching a peak in November. The highest sales occur in November,impact of holiday shopping, suggesting a strong seasonal effect.December also maintains elevated sales, reinforcing the holiday-driven demand pattern.The following charts illustrate clear seasonal patterns and trends in monthly sales performance.

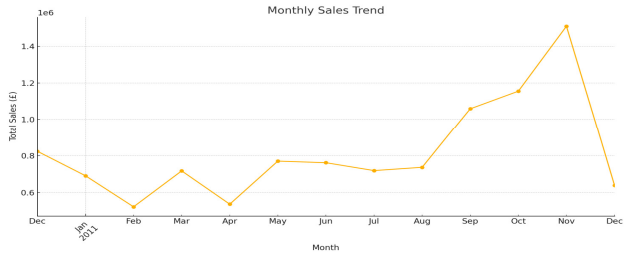

Figure 1: Monthly Sales Trend Line Chart

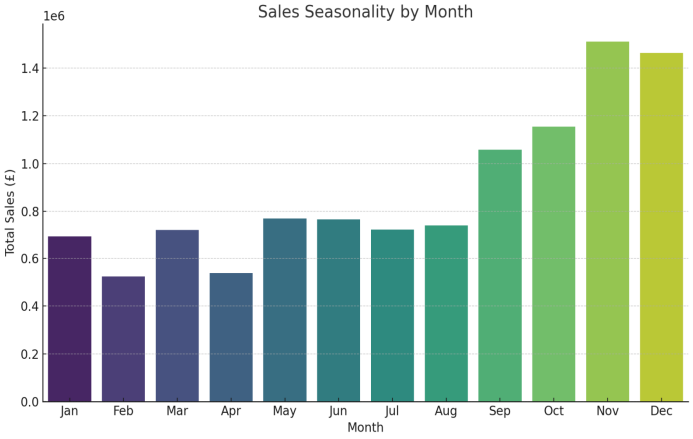

Figure 2: Sales Seasonality by Month

4.2 Product Performance

The analysis of top-selling products reveals key insights into customer preferences and revenue drivers. The top-selling product is *dotcom postage* suggesting a significant portion of revenue is related to shipping or online orders. *Regency cakestand 3 tier* is also a major contributor, possibly indicating high demand for decorative or entertaining products.H*anging heart t-light holder*, *party bunting*, and *jumbo bag red retrospot* all generate substantial revenue. Items like *rabbit night light*, *paper chain kit 50's christmas*, and *assorted colour bird ornament* show that seasonal and decorative products contribute meaningfully. Top Selling products by revenue are shown in figure 3,

The Popular products are represented in product description word cloud (Figure 4). It visualizes the most frequent words found in product descriptions, with larger and bolder words appearing more frequently.This tool is particularly useful for SEO and keyword optimization, helping businesses tailor product listings for better visibility and search performance.

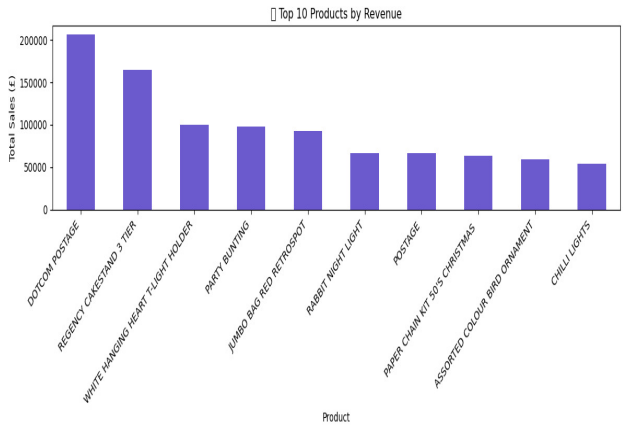

Figure 3: Top 10 Products by Revenue (Bar Chart)

Figure 4: Product keyword Word Cloud

4.3 Sales by Geography

The United Kingdom overwhelmingly leads with over 85% of total revenue, followed by Netherlands, Ireland, and Germany. This confirms the UK as the primary market, while also highlighting potential for growth in nearby European countries. Figure 5 presents a chart illustrating total sales by country, highlighting the distribution of revenue across different geographies

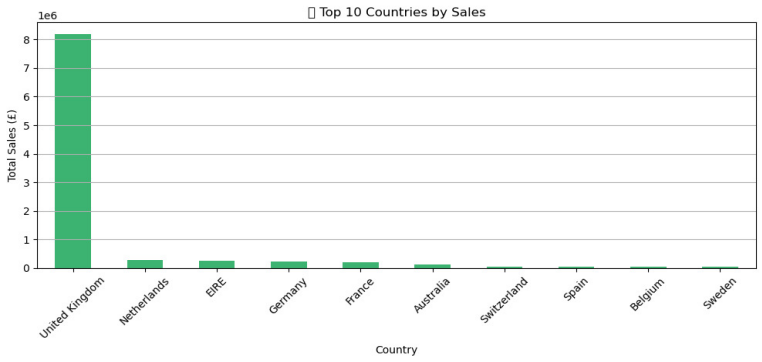

Figure 5: Sales by Country (Bar Chart)

4.4 Correlation Analysis:

The correlation analysis includes a heatmap (Figure 6) that visualizes the relationships between key numerical variables—Quantity, Unit Price, and Total Sales.

There is a strong positive correlation (0.89) between Quantity and Total Sales, indicating that sales are primarily driven by volume. In other words, higher quantities sold directly lead to higher total sales.

There is a weak negative correlation (-0.16) between Unit Price and Total Sales, suggesting that price changes alone have a limited direct effect on overall sales performance.

Additionally, the correlation between Quantity and Unit Price is approximately zero, indicating no significant relationship. This implies that the quantity sold is largely independent of the product's unit price.

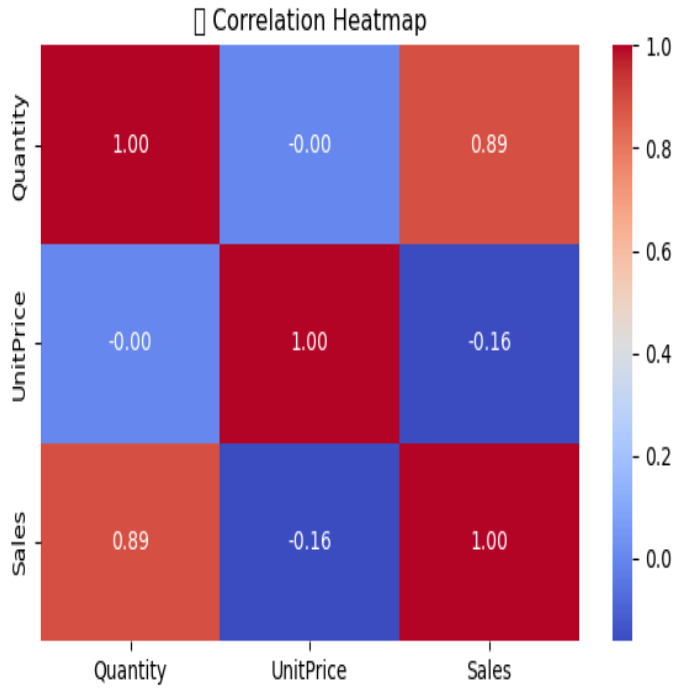

Figure 6: Sales Correlation Heatmap

4.5 Customer Segmentation

RFM (Recency, Frequency, Monetary) analysis was used to evaluate customer behavior. Using K-Means clustering, customers were grouped into clusters.Cluster summary and segmentation chart have been given in Table 1 and figure 7 respectively.

Table 1 -Cluster Summary

	Recency	Frequency	Monetary	Count
Cluster				
0	74.1	2.9	1089.0	1851
1	575.5	1.1	318.4	151
2	46.9	21.6	24010.3	56
3	254.6	1.4	497.4	939

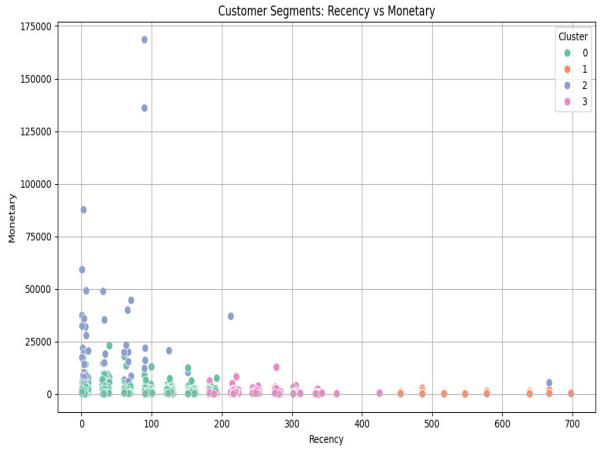

Fig 7: Customer Segmentation Scatter Plot

Interpretation of Scatter Plot :

- X-axis = Recency (lower is better, more recent)
- Y-axis = Monetary (higher is better)
- Clusters Details -

- Cluster 2 - Top left VIPs (recent, frequent, big spenders)

- Cluster 0 - Mid left= Regular customers (recent, but moderate purchases).

- Cluster 1 -Bottom Right= Lost or likely churned (long time since last purchase).

- Cluster 3 = Low-value or occasional ,at risk (somewhat old, low purchases).

Based on customer segmentation insights, targeted strategies can be applied to each cluster: high spenders should be rewarded and upsold, churned and at-risk groups need re-engagement efforts, and average customers can be incentivized to increase purchase frequency through personalized offers.

4.6 Predictive Modeling- Sales Forecasting

In predicting modeling , the focus was on predicting future sales to aid inventory planning, budgeting, staffing, and resource allocation. We built and evaluated three models: ARIMA (AutoRegressive Integrated Moving Average),

Linear Regression and K-Nearest Neighbors (KNN). Using relevant evaluation metrics, we compared the models' performance, interpreted the results, and recommended the best model for accurate sales forecasting. The findings are presented in Table 2 and visualized in Figure 8.

Table 2: Model Performance Comparison

Model	MAE	RMSE	R2
Linear Regression	11610.05	15047.15	0.59
KNN	12304.05	15903.45	0.54
ARIMA	24737.17	28705.50	0.50

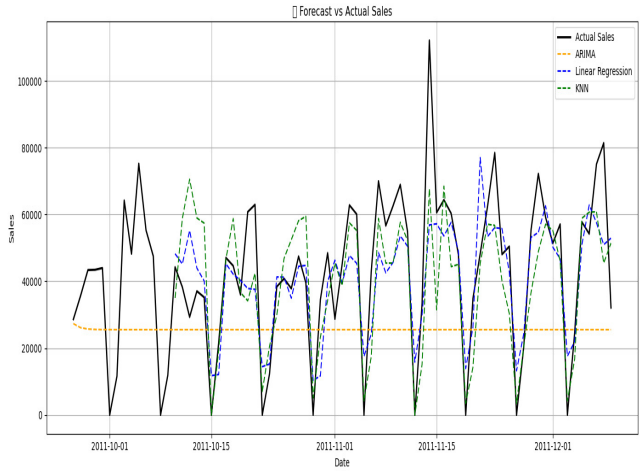

Figure 8: Sales Forecasting comparison:ARIMA, Linear Regression, and KNN

The results from the performance metrics show that Machine Learning models (Linear Regression and KNN) are better than the Time Series model (ARIMA) for this forecasting task. They offer more accurate and reliable predictions with lower errors, making them a preferable choice for this specific problem.Linear Regression performs the best overall, with the lowest MAE and RMSE. KNN follows closely behind, with slightly higher MAE, RMSE.

5. Business Insights

The data-driven analysis revealed several key insights for optimization.

- **Seasonality is strong**, especially around December, indicating a need for promotional and inventory planning during Q4.
- **Top-selling products** have themes around home decor and gifting — guiding procurement and marketing focus.
- The **UK dominates revenue**, suggesting strong brand penetration locally but opportunities for growth abroad.
- **Linear Regression model outperformed** other forecasting models, offering the most accurate sales predictions.
- **Customer Segmentation** revealed a small cluster of highly profitable, loyal buyers - ideal for loyalty programs and retention strategies. Strategic actions should be tailored to each customer segment: premium offers and exclusive perks for high-value loyal customers, engagement campaigns for moderate-value groups, and reactivation or acquisition strategies for low-engagement segments.

6. Conclusion

In an increasingly data-centric retail environment, data-driven analytics has become a cornerstone of modern business intelligence, empowering organizations to extract actionable insights from large volumes of transactional data. This study underscores the value of integrating predictive analytics and customer segmentation techniques to enhance sales forecasting and strategic decision-making in the online retail sector.

By employing both classical time series models (such as ARIMA) and machine learning approaches (including Linear Regression and K-Nearest Neighbors), the analysis demonstrated that machine learning models offer superior accuracy in forecasting sales. Furthermore, the use of RFM analysis and K-Means clustering enabled effective customer segmentation, revealing a high-value customer segment suitable for targeted loyalty and retention strategies.

The comprehensive study presented in this paper provides a practical roadmap for optimizing inventory management, marketing efforts, and customer engagement. It supports data-informed decision-making and offers valuable insights into how personalized strategies can drive profitability in competitive e-commerce markets.

Future research may explore enhancements such as real-time forecasting, advanced feature engineering, customer churn prediction, and integration with recommendation systems to further refine predictive performance and strategic personalization.

REFERENCES

1. Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning–A comparative analysis. *Sustainable Futures*. https://www.sciencedirect.com/science/article/pii/S2667096822000027
2. Sharma, A., Patel, N., & Gupta, R. (2021). Enhancing retail sales forecasting through LSTM networks and ARIMA models: A comparative analysis of AI methodologies. *European Advances and Applications in Artificial Intelligence Journal*. Retrieved from http://www.eaaij.com/index.php/eaaij/article/view/7
3. Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data*. Retrieved from https://www.mdpi.com/2306-5729/4/1/15
4. Gaytri Devi & Sukhija.V ,Transforming missing data into valuable insights:A computational intelligence perspective,international journal of all research education and scientific methods (IJARESM),ISSN: 2455-6211, Volume 11, Issue 9, September-2023,
5. Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University – Computer and Information Sciences*. Retrieved from https://www.sciencedirect.com/science/article/pii/S1319157819309802
6. Hallimsyore Kalegowda, A. (2023). Utilizing Predictive Analytics to Enhance Retail Business Performance. *National College of Ireland Thesis Repository*. Retrieved from https://norma.ncirl.ie/7523/
7. Kadam, V., & Vhatkar, S. (2022). Design and develop data analysis and *Data Mining and Information Security* (pp. 165-177). Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-981-16-4863-2_14

