

Sales Data Analysis and Future Sales Prediction Using Machine Learning: An Empirical Study on E-Commerce Dynamics and Demand Forecasting

Praveenraj R, Ashwin Kumar S, Suryaprakash S N

CSDA, Dr.N.G.P. Arts and Science College, Coimbatore, India CSDA, Dr.N.G.P. Arts and Science College, Coimbatore, India
262DA017.cs@drngpasc.ac.in,262DA002. cs@drngpasc.ac.in,262DA029.cs@drngpasc.ac.in

Abstract

Reliable sales forecasting and detailed transaction analysis are important for e-commerce businesses because they help teams understand customer demand and plan daily operations. This study examines the Online Retail transaction dataset, covering the period from 1 December 2010 to 9 December 2011. The analysis applies a series of data-cleaning steps to remove cancelled invoices, non-positive prices and quantities, and duplicate records. Valid transactions are retained, including genuine purchases made by customers without an identified customer number. The cleaned data records total sales of GBP 10,642,110.80 across 305 active trading days. To evaluate forecasting performance fairly, the data is split chronologically into 80% for training and 20% for testing, helping prevent information from the future from influencing the models. Two simple forecasting methods—naive lag-1 persistence and a 7-day moving average—are compared with Ridge Regression and Random Forest. Of the four approaches, Random Forest performs best, with a Mean Absolute Error (MAE) of GBP 16,328.43 and a Root Mean Squared Error (RMSE) of GBP 26,595.74. Its MAE is 25.81% lower than that of the naive baseline. The findings also highlight seasonal changes in sales, the absence of Saturday transactions, and the increase in demand before Christmas. These patterns can support better inventory and purchasing decisions.

Keywords: Sales forecasting, e-commerce analytics, machine learning, Random Forest, time-series analysis, data leakage prevention, and retail demand planning.

1. Introduction

Online shopping has grown rapidly over the past two decades. As a result, e-commerce businesses now collect large volumes of transaction data that reveal purchasing habits, product preferences, and changes in demand across markets. Turning this raw information into reliable sales forecasts, however, is not always straightforward. Retail transaction data can contain unusual values and patterns, including cancelled orders, postage charges, free promotional items, and inventory adjustments. Sales can also change from one season to another, making it difficult to identify trends using simple methods alone.

Sales forecasts influence several important business decisions, such as when to restock products, how to manage cash flow, how to organize deliveries, and how many warehouse employees are needed. If demand is underestimated, businesses may run out of stock and lose sales. If it is overestimated, they may hold excess inventory and face additional storage costs or discount products to clear them. This study develops and evaluates a repeatable approach to cleaning transaction data, exploring sales patterns, and forecasting future revenue with

machine-learning models.

2. Literature Review and Related Work

Retail sales forecasting has traditionally used statistical methods such as Autoregressive Integrated Moving Average (ARIMA) and exponential smoothing (Hyndman & Athanasopoulos, 2018). These methods are useful because their assumptions and results can often be explained clearly. However, they may have difficulty capturing complicated non-linear relationships, holiday-related sales increases, and changes in trading schedules unless the relevant features are carefully designed.

Supervised machine-learning methods offer another way to model sales data. Regularized linear models, including Ridge Regression (Hoerl & Kennard, 1970), and tree-based ensemble methods such as Random Forest (Breiman, 2001) can capture a wider range of relationships in structured data. Random Forest and gradient-boosted trees can also learn interactions between past sales values and changing patterns without requiring the data to follow strict stationarity assumptions. However, a model must be tested carefully: as discussed in forecasting research, the training data should come before the test data in time. This chronological approach helps prevent information from future observations leaking into the model (Tashman, 2000).

3. Data Hygiene and Preprocessing Methodology

This study uses the Online Retail transaction dataset available through the UCI Machine Learning Repository (Chen et al., 2012). It contains purchases made through a UK-based online giftware and homeware wholesaler between 1 December 2010 and 9 December 2011. The initial dataset contains 541,909 transaction records.

We cleaned the data in several stages to improve its quality and reduce the risk of overstating sales. The steps and the number of records remaining after each stage are shown in the data-cleaning table.

Step	Hygiene Stage	Remaining Records	Decision Rationale
0	Raw Ingestion	541,909	Base ingestion of UCI Online Retail dataset.
1	Remove Cancellations	532,621	Invoices prefixed with 'C' indicate cancelled/reversed orders; excluded from positive sales revenue.
2	Filter Non-Positive Quantities	531,285	Quantities ≤ 0 represent broken items, inventory losses, or return adjustments; excluded from gross sales.

3	Filter	Non-Positive	530,104	UnitPrice <= 0 represents promotional zero-cost items, manual bookkeeping, or debt write-offs; excluded from revenue analysis.
Unit		Prices		
4		Deduplication	524,878	Identical duplicate transactions removed to prevent artificial inflation of sales volume.
5	Handle	Missing	524,878	Retained 132,186 guest checkouts with CustomerID=0 to preserve genuine commercial revenue; imputed 0 missing descriptions.
Metadata				

After the cleaning process, the final analytical dataset contains 524,878 valid commercial transactions, retaining 96.86% of the original records. Total sales revenue is calculated for each transaction using the following formula:
 $SalesRevenue_i = Quantity_i * UnitPrice_i$
 Using this calculation, total sales revenue across the cleaned dataset is GBP 10,642,110.80.

4. Exploratory Data Analysis & Empirical Patterns

We grouped the transaction data by time and location to understand how sales changed during the study period. Table 2 summarizes monthly revenue, order counts, the number of items sold, and average order value.

Year-Month	Total Revenue (GBP)	Orders	Items Sold	Avg Order Value (GBP)
2010-12	£821,452.73	1,559	358,019	£526.91
2011-01	£689,811.61	1,086	387,099	£635.19
2011-02	£522,545.56	1,100	282,934	£475.04
2011-03	£716,215.26	1,454	376,599	£492.58
2011-04	£536,968.49	1,246	307,953	£430.95

2011-05	£769,296.61	1,681	395,001	£457.64
2011-06	£760,547.01	1,533	388,511	£496.12
2011-07	£718,076.12	1,475	399,693	£486.83
2011-08	£757,841.38	1,361	421,020	£556.83
2011-09	£1,056,435.19	1,837	569,573	£575.09
2011-10	£1,151,263.73	2,040	621,029	£564.34
2011-11	£1,503,866.78	2,769	751,377	£543.11
2011-12	£637,790.33	819	313,612	£778.74

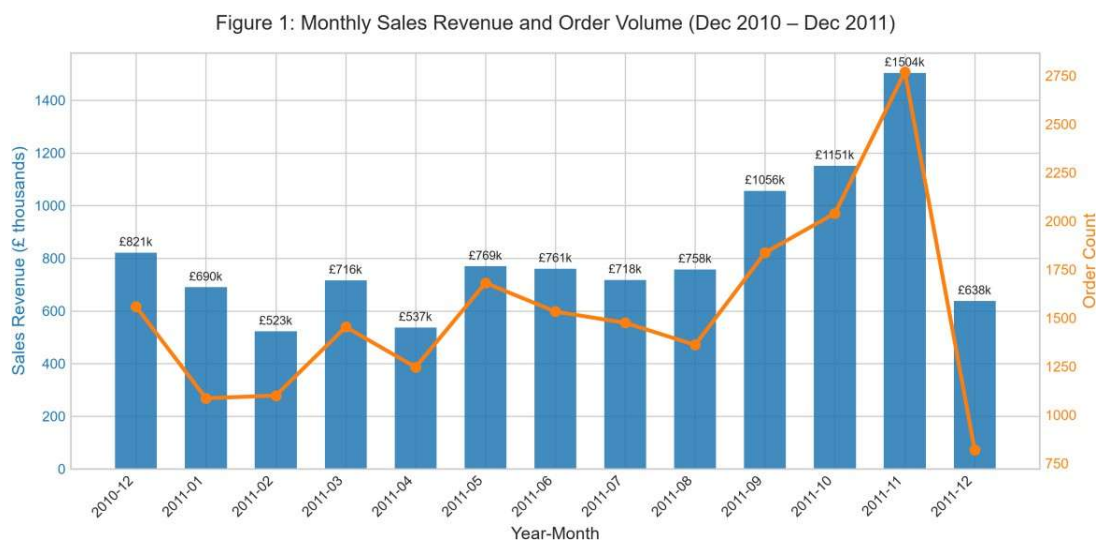


Figure 1. Monthly sales revenue and order volume from December 2010 to December 2011.

Figure 1 shows that monthly revenue generally increased during 2011, reaching its highest point in November. Sales that month totalled GBP 1,503,866.78 across 2,769 orders, with 751,377 items sold. This increase may reflect retailers purchasing stock ahead of the Christmas season. December revenue was lower at GBP 637,790.33, but the dataset ends on 9 December, so it does not represent the full month.

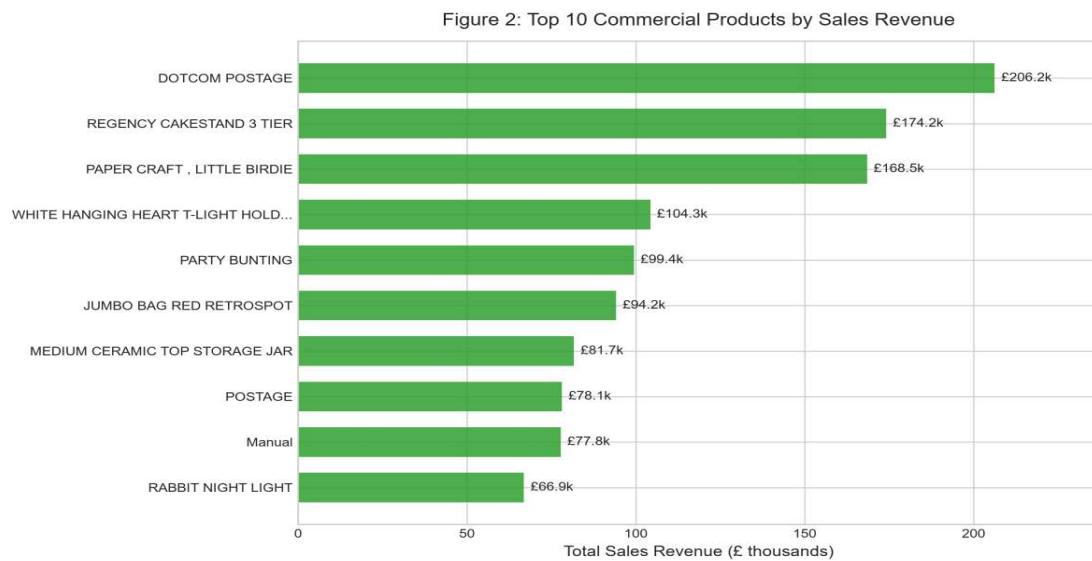


Figure 2. Ten products with the highest cumulative sales revenue.

The product-level results show that a relatively small group of products contributed substantial revenue. Excluding postage, the leading products included 'REGENCY CAKESTAND 3 TIER' (GBP 174,156.54), 'PAPER CRAFT, LITTLE BIRDIE' (GBP 168,469.60), and 'WHITE HANGING HEART T-LIGHT HOLDER' (GBP 104,284.24).

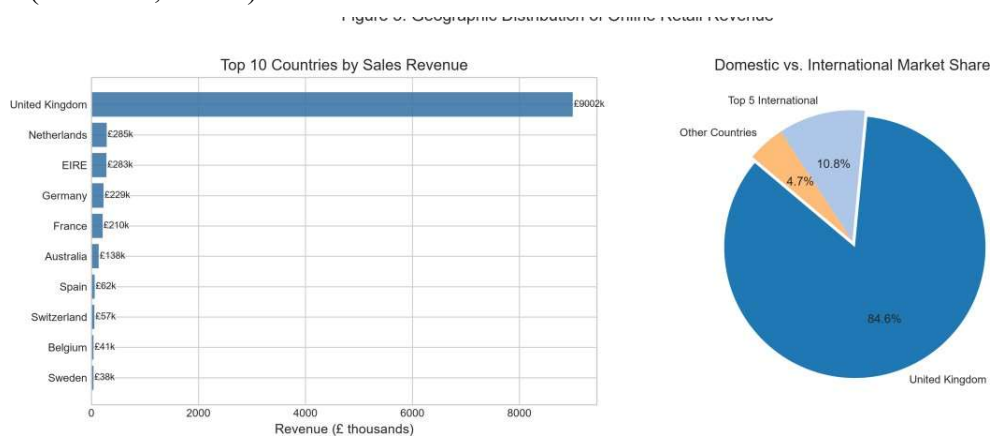


Figure 3. Sales revenue across the United Kingdom and international markets.

The United Kingdom accounted for GBP 9,001,744.09, or 84.58% of total revenue. The main international markets included the Netherlands (GBP 285,446.34), EIRE (GBP 283,140.52), Germany (GBP 228,678.40), and France (GBP 209,625.37). This suggests that the business relied mainly on its domestic market while also serving customers across Western Europe.

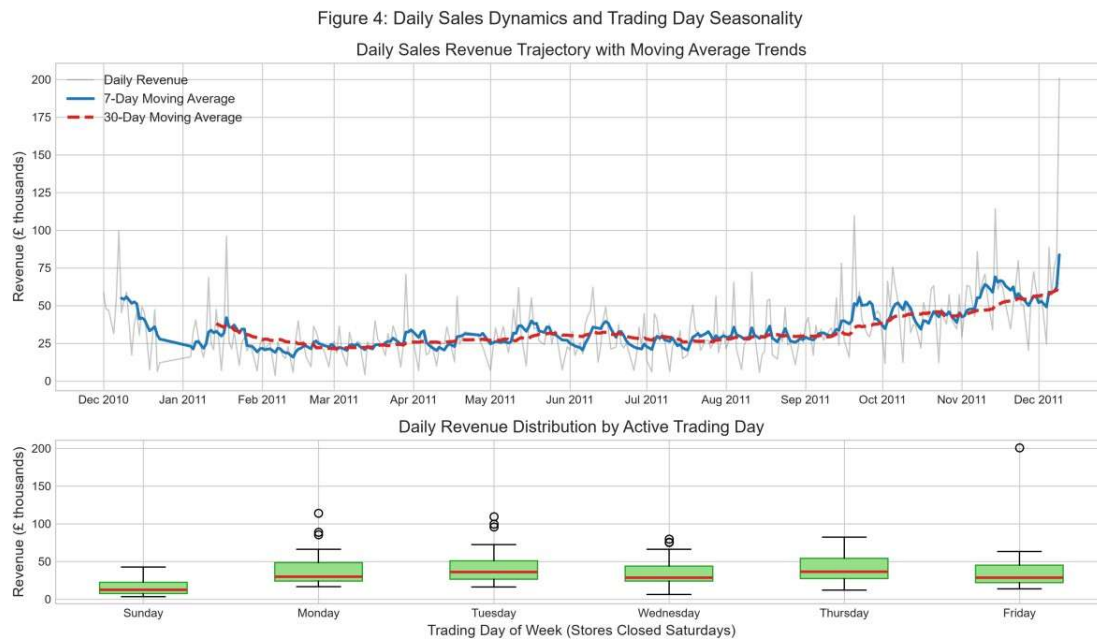


Figure 4. Daily sales revenue patterns and differences across trading days.

The dataset records sales on 305 active trading days. No transactions were recorded on Saturdays during the 373-day period, which is consistent with the wholesaler being closed on that day. Average daily revenue was GBP 34,892.17, while Wednesday and Thursday had the highest median revenue among the trading days.

5. Predictive Modeling and Experimental Setup

We treated daily sales forecasting as a supervised learning problem. In this approach, the model estimates revenue for a given day using information available from earlier days only. The input features included the following:

1. Past sales values (lags): revenue from 1, 2, 3, and 7 days earlier.
2. Rolling statistics: the 7-day rolling average and standard deviation, and the 14-day rolling average. These measures were shifted by one day so that they did not include the revenue being predicted.
3. Calendar features: day of the week, day of the month, month, a Sunday indicator, and a linear time-trend index.

To prevent data leakage, we split the 291 days with valid features into training and testing sets in chronological order, after allowing for a 14-day warm-up period. The training set contains 232 days from December 2010 to October 2011, while the test set contains 59 days from October to December 2011. We did not shuffle the dates or use information from the test period when preparing the training data.

We compared four forecasting approaches:

1. Naive baseline (lag-1 persistence): predicts that the next day's revenue will be the same as the previous day's revenue.
2. 7-day moving-average baseline: predicts revenue using the average of the previous seven days.

3. Ridge Regression: a linear model with L2 regularization ($\alpha = 10.0$) and standardized features.
- 6.Random Forest Regressor: a non-linear ensemble model with 100 trees and a maximum tree depth of
- Experimental Results & Performance Analysis

We evaluated all four models over the same 59-day test period. The measures used were Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). The results are summarized in Table 3.

Model	MAE (GBP)	RMSE (GBP)	MAPE (%)	R^2 Score	MAE Gain vs Naive (%)
Naive Persistence (Lag-1)	£22,009.64	£30,368.30	50.09%	-0.2975	0.00%
Moving Average (7-Day)	£16,867.56	£26,194.00	38.52%	0.0347	23.36%
Ridge Regression	£16,725.93	£27,307.63	30.31%	-0.0491	24.01%
Random Forest Regressor	£16,328.43	£26,595.74	27.04%	0.0048	25.81%

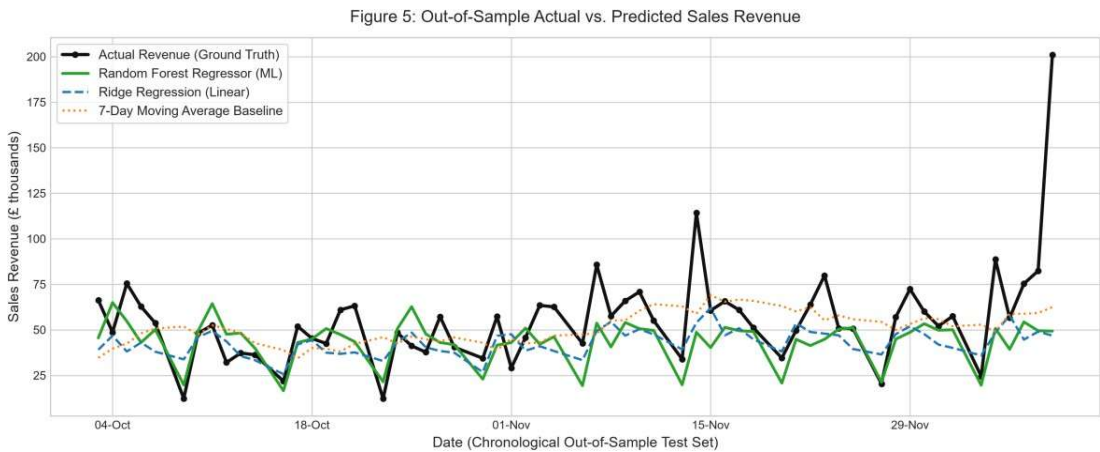


Figure 5. Actual daily sales revenue compared with predictions from each model on the test period.

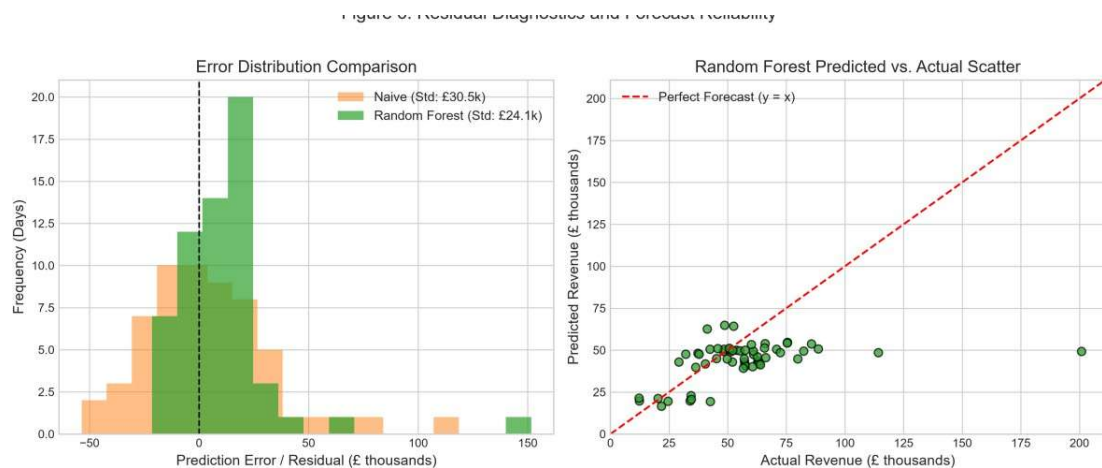


Figure 6. Comparison of prediction errors and a scatter plot of actual versus predicted revenue.

The main findings from the model comparison are as follows:

1. Random Forest produced the lowest MAE (GBP 16,328.43) and an RMSE of GBP 26,595.74. Its MAE was 25.81% lower than the naive baseline's MAE of GBP 22,009.64, making it the strongest model among those tested.
2. Ridge Regression also performed reasonably well, with an MAE of GBP 16,725.93. This suggests that past sales and rolling statistics provide useful information even for a linear model.
3. The 7-day moving average improved on the naive baseline, reducing MAE from GBP 22,009.64 to GBP 16,867.56. Averaging the previous week's revenue appears to reduce some of the day-to-day variation in sales.
4. Figure 6 compares the prediction errors. The Random Forest errors are more concentrated around zero than those from the naive baseline, although the scatter plot also shows that some daily sales values remain difficult to predict accurately.

7. Managerial Discussion and Implications

In practical terms, daily sales forecasts can help retailers plan purchasing and manage the money tied up in stock. A Random Forest model that uses past sales, rolling averages, and calendar information could help purchasing teams plan supplier orders one to two weeks ahead and reduce sudden changes in ordering across the supply chain, often called the 'bullwhip effect'. Understanding how demand varies around the weekend may also help warehouse managers plan staffing and prepare for orders that accumulate when operations are closed. These uses should be assessed in the context of the business's actual operating schedule.

8. Limitations and Threats to Validity

The results should be considered alongside several limitations:

1. Single-business dataset: the data comes from one UK giftware wholesaler. The results may not apply directly to other industries, such as grocery retail or fast fashion, without further testing.
2. Missing customer identifiers: around 25% of transactions did not have customer IDs. These transactions were retained for revenue analysis, but they cannot support detailed tracking of individual customers over time.

3. Incomplete holiday period: the dataset ends on 9 December 2011, so it does not include the full Christmas period or post-Christmas and New Year clearance sales.

4. Missing external factors: the dataset does not include information about wider economic conditions, competitors' prices, or digital advertising spending, all of which may influence sales.

9. Conclusion and Future Research

This study developed a complete workflow for analysing e-commerce sales and forecasting future revenue. It combined data cleaning, features based only on past information, and a comparison of baseline and machine-learning models. On the test data, Random Forest achieved the lowest MAE among the four approaches evaluated. The results suggest that machine learning can support sales planning, although the findings are based on one retailer and a limited time period. Future work could compare the current approach with deep-learning models such as Temporal Fusion Transformers and N-BEATS, and explore forecasts at individual product (SKU) level while ensuring that product-level predictions remain consistent with overall sales totals.

References

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3), 197-208. <https://doi.org/10.1057/dbm.2012.17>
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67. <https://doi.org/10.1080/00401706.1970.10488634>
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts: Melbourne, Australia.
- Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4), 437-450. [https://doi.org/10.1016/S0169-2070\(00\)00078-4](https://doi.org/10.1016/S0169-2070(00)00078-4)