# Predicting online retail demand: Machine Learning Approach

Eduardo Capel Alguacil
*Departament of Computer Science*
*City, University of London*
London, United Kingdom
eduardo.capel-alguacil@city.ac.uk

***Abstract*— The changing environment has led to challenges in the retail sector, particularly in managing the delicate equilibrium between supply, demand, and customer preferences. Retailers face the challenge of managing inventory levels and trying to understand their customers purchasing behaviors. Stock-outs and excess of stock levels lead to significant financial implications and therefore the issues become a crucial matter in the retail sector.**

**This study leverages data from UCI Machine Learning repository, analyzing transactions of customers from 2011 to enhance demand forecasting, evaluating variable effects on demand, and trying to find hidden customer behaviors within the data. Revealing insights of buying patterns and demographics, can guide retailers to adapt their strategies to meet evolving market demands. It also assesses the effectiveness of linear regression and XGBoost models to forecast future demand, provide actionable solutions for the retail market to navigate through different challenges.**

***Keywords—retail, demand, forecasting, clustering, inventory management.***

## I. INTRODUCTION

The retail industry faces a variety of obstacles as a result of constantly changing environments in which it operates. The constant battle with supply and demand imbalances is one of these difficulties, made worse by the extreme variety of consumer tastes. Retailers of today operate in a market characterized by social tastes that are unprecedentedly diverse, marked by fierce competition and complexity that reflects the many facets of customer needs [2].

Customers become important players in determining how retail businesses do financially in this complex environment. However, the effect isn't limited to what customers do, in the retail industry, disparities between inventory levels and customer desires have a big impact. Stock-outs have immediate negative effects, as they result in lost sales opportunities. In addition to this, an excess in the amount of product in stock creates inefficient operations and sometimes even spoilage. These discrepancies between stock and customer needs have dramatic financial consequences. Usually, the losses of lost sales are greater than the costs of keeping excess inventory on hand. Retailers are being driven by this economic reality to fulfill a crucial strategic requirement, which is to maintain a service level that exceeds 50% for each specific product. This dedication to make sure products remain in stock is indicative of a proactive strategy meant to prevent revenue losses from lost sales [1].

It takes a balancing act to navigate this retail world, and demand forecasting becomes essential to efficient inventory management. The integration of forecasting in inventory management, as seen in companies like Amazon, results in significant improvements [5] reducing overstock and out of stock scenarios within the business. Therefore, businesses should make concerted efforts to forecast demand.

## II. ANALYTICAL QUESTIONS

Our study uses data belonging to UCI Machine Learning repository, containing transactions made in the year 2011 by UK-based and non-UK-based customers in the retail sector. It is commonly used in different research studies, which makes it suitable for generating meaningful and actionable insights in our study. It perfectly aligns with addressing our research questions.

As mentioned before, stock-outs and the diversity in customer tastes play a pivotal role in the business retail sector since this can provoke losses in the short and long term for the organization. Inventory management can be leveraged by segmenting customer based on their purchase behavior, leading to more efficient practices.

Hence, it is logical to assume that businesses in the retail sector should also pay attention to avoid losses caused by bad management of inventory. While identifying different clusters of customers is informative, our goal is to predict and understand the demand. In order to achieve this demand understanding, customers segments behavior and foremost, customer characteristics play a crucial role in this process. To give our study a clear direction, the following are our research inquiries.

*What insights can be drawn from seasonal sales trends, and how can these insights inform strategic business decisions?*

*How do variations in pricing influence the predictability of monthly sales demand?*

*Could clustering help uncover hidden customer segments within the dataset, and how do these segments relate to sales and price?*

*What is the significance and practical impact of the differences in sales patterns among customer clusters compared to the overall market?*

*How effectively do linear regression and XGBoost models predict future sales based on historical data, especially considering the incorporation of lag variables?*

## III. DATA

The data presented contains 541.909 observations and 8 variables, each playing a distinct role in understanding and predicting sales patterns. Variables like 'InvoiceNo' and 'StockCode' uniquely identify transactions and products, while 'Quantity' and 'UnitPrice' provide essential insights into purchasing behaviors. 'CustomerID' and 'Country' contribute to the demographic understanding of consumers, enabling targeted strategies. Lastly, 'InvoiceDate' serves as a temporal variable, facilitating trend analysis and temporal correlation of sales data. To align the data with the main study's objectives, predicting monthly demand, a strategic grouping of variables is employed, synthesizing information based on customer IDs, product IDs, descriptions, countries, and monthly invoice dates. This aggregation strategy allows for a comprehensive analysis, including summation of quantities sold, mean unit prices, counting product variations, and identifying unique sales days for enriched insights.

Table I

| | CustomerID | Product_ID | Description | Country | InvoiceDate | Monthly_sales | Monthly_price | Total_transactions | Unique_days_with_sales |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 12346.0 | 23166 | MEDIUM CERAMIC TOP STORAGE JAR | United Kingdom | 2011-01-31 | 74215 | 1.04 | 1 | 1 |
| 1 | 12347.0 | 16008 | SMALL FOLDING SCISSOR(POINTED EDGE) | Iceland | 2011-04-30 | 24 | 0.25 | 1 | 1 |
| 2 | 12347.0 | 17021 | NAMASTE SWAGAT INCENSE | Iceland | 2011-06-30 | 36 | 0.30 | 1 | 1 |
| 3 | 12347.0 | 20685 | RED RETROSPOT PURSE | Iceland | 2011-04-30 | 6 | 2.95 | 1 | 1 |
| 4 | 12347.0 | 20719 | WOODLAND CHARLOTTE BAG | Iceland | 2011-01-31 | 10 | 0.85 | 1 | 1 |

The dataset has undergone careful preprocessing, including encoding for customer and country attributes, to make this analysis possible. The collection also includes a number of important variables that capture spending patterns, consumer transactional patterns, temporal features, and frequency-encoded representations. Together, these factors serve as the foundation for extensive research that aims to identify hidden client groups and distinguish particular behavioral patterns in pricing and sales preferences, aiming to provide insight into customer segments with distinct market behaviors.

Table II

| CustomerID | 12347.0 | 12347.0 | 12347.0 | 12347.0 | 12347.0 |
|---|---|---|---|---|---|
| Total_transactions | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Unique_days_with_sales | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| TotalSpending_avg | 23.681319 | 23.681319 | 23.681319 | 23.681319 | 23.681319 |
| InvoiceDate_month | 8.000000 | 10.000000 | 12.000000 | 4.000000 | 12.000000 |
| CustomerID_FrequencyEncoded | 0.000501 | 0.000501 | 0.000501 | 0.000501 | 0.000501 |
| country_FrequencyEncoded | 0.000501 | 0.000501 | 0.000501 | 0.000501 | 0.000501 |
| Log_Monthly_sales | 2.397895 | 2.397895 | 2.397895 | 1.945910 | 3.218876 |
| Log_Monthly_price | 0.615186 | 0.615186 | 0.615186 | 1.373716 | 0.810930 |

Following a similar strategy, predictive models aim to predict the monthly demand. The data has been slightly transformed from the previous task, creating variables such as the lag sales variable, since future sales can be influenced by past sales trends[3].

Table III

| | 84923 | 207151 | 362429 | 362428 | 269772 |
|---|---|---|---|---|---|
| Total_transactions | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Unique_days_with_sales | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| TotalSpending_avg | 7.496355 | 17.243155 | 19.873684 | 19.873684 | 20.725766 |
| InvoiceDate_month | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| ProductID_FrequencyEncoded | 0.000349 | 0.000132 | 0.001282 | 0.001700 | 0.000140 |
| CustomerID_FrequencyEncoded | 0.000558 | 0.001081 | 0.000366 | 0.000366 | 0.000305 |
| country_FrequencyEncoded | 0.888433 | 0.888433 | 0.888433 | 0.888433 | 0.888433 |
| Log_Monthly_sales | 0.693147 | 3.891820 | 3.218876 | 3.218876 | 1.945910 |
| Log_Monthly_price | 2.393339 | 0.350657 | 0.438255 | 0.438255 | 1.658228 |
| Monthly_logsales_lag | 2.197225 | 2.197225 | 3.218876 | 3.218876 | 1.945910 |

## IV. ANALYSIS

### A. *Data aggregation*

In order to cope with the overall goal of this study, which is forecasting the monthly demand, the initial dataset has undergone different aggregations. For these aggregations, customers, products, descriptions, and Country were used as attributes. Furthermore, the transaction invoice served as temporal variable to make aggregations month by month, this means that all transactions, made by a singular customer, for a specific product in a specific country with and specific description are all aggregated by months.

For these groups, different calculations are performed, total quantity of products sold is summed for each group, and the average unit price is calculated for each group as well. Moreover, the count of total transactions reflects how frequently a product was bought and the unique days with sales indicate the spread of sales through different days. Other relevant variables such as the customer average spending and lag sales were calculated.

### B. *Data preprocessing*

In the data preprocessing stage, the data goes through two main transformations. The first one is related to the transformation of the categorical data, product, customer, and country variables. To make them suitable for inputting them in the model, frequency encoding is the choice of this transformation, computing the frequency of occurrences within the data set. This choice of categorical variables preprocessing is due to the number of distinct customers, products, and countries, since the creation of dummy variables excessively increases the number of variables in the data. In addition, the frequency of occurrences is normalized facilitating the exploration of potential groups within the data.

The second transformation is related to continuous variables, owing to the extreme positive skewness observed in both monthly sales and monthly prices. Log transformation is applied to both variables summing up 1 as a constant. This transformation mitigates skewness, leading to a more normalized shape of both variables, helping to stabilize variance across retail dataset.

### C. *Explanatory data analysis*

The explanatory analysis aims to create insight particularly valuable in terms of strategic business decisions. This analysis shed light on sales and income behavior throughout the 2011 year, revealing patterns, and months peak season in terms of sales and income. This can guide the business to optimize different strategies related to sales, inventory management and, promotions to capitalize on peak buying months.

Secondly, the average spending in each country is explored, this average spending is distributed across different countries where the retail market operates, offering a better understanding of the business market and customers preferences. This can be used from a business perspective to

target marketing strategies, efficiently allocate resources, and manage inventory.

### *D. Modelling*

In the first modelling part, it aims to study the correlation between the target variable, monthly sales with the rest of variables, displaying how positively or negatively are variables correlated to monthly sales and considering the nature of variables, since some variables such as lag sales and revenue contain the target in them and therefore, they show higher correlation results.

Most correlated variables served to highlight the importance of this variable in building a model. To study in detail their effect in sales variable, a linear regression model is constructed, observing how much variability in sales can be explained using the variable price. Through this, the model provides insights in terms of key drives in sales performance and can lead to develop more accurate forecasting models.

In addition to this, seeking to find hidden groups of customers who have different behaviors in this retail market, K-means clustering has been conducted using different variables. These clusters aim to capture various aspects of customer behavior, looking for their potential use in building the demand forecasting model. It can also lead to the use of actionable marketing strategies and customer engagement initiatives.

Using elbow plot, the number of clusters is determined and using previous analysis, target variable monthly sales and the most correlated variable, monthly prices serve to explore the cluster behavior, looking how they distribute within these two variables, seeking to find different behaviors. Similarly, using these two variables, different linear regression, and polynomial models of order 2 were constructed represented by each cluster. This aims to reveal noticeable trends of clusters and look for unique patterns and behavior identified. Overlay distributions and models can lead to the unsuccessful identification of unique clusters behaviors.

To understand the significance of cluster in monthly sales, a T-test is conducted, examining how their means sales of each cluster differ from overall sales in this market. This can indicate distinct sales patterns in each cluster. To further study cluster effect on monthly sales, Cohen's distance is used to measure the real effect clusters have in monthly sales. The cluster effect size indicates impact of clusters on monthly sales and can lead to a better understanding of how clusters drive sales. A small size effect implies lower impact and larger effect a more pronounce influence on sales.

### *E. Forecasting*

The last part of this analysis aims to compare a linear regression model and a XGBoost model forecasting the future sales demand. Models are trained using early data and tested on the more recent data, specifically in the last 2 months of 2011. To allow comparison of results, R-squared and Root Mean Squared Error metrics are used. This comparative analysis allows us to determine which model is more efficient in capturing the complexity of the data and predicting future sales demand. Predictors variables used in both models are also explored, analyzing variable weights, and seeking to find the contributors to sales demands and how different variables affect the forecasting accuracy.

### *F. Validation of results*

As mentioned previously, R-squared and Root Mean Squared Error (RMSE) are used to validate results. The coefficient of determination represents the proportion of variance in the dependent variable that is predictable from the independent variable. RMSE measures the differences between predicted values and the observed values [4].

$$\frac{\Sigma(yi - \hat{y}i)^2}{\Sigma(yi - \bar{y})^2}$$

R-squared

$$\sqrt{(\Sigma(yi - \hat{y}i)^2 / n)}$$

Root Mean Squared Error´

## V. RESULTS

Different insights and results can be obtained from the previous analysis made. Regarding the seasonal sales trends analysis, it can be observed that there is an increase in monthly sales as well as an increase in income, progressively for the 2011 year. The data revealed pronounced sales during specific periods, specifically during September, October, and November. It can be observed that most sales, expressed in total spending by each country, heavily rely on the United Kingdom.

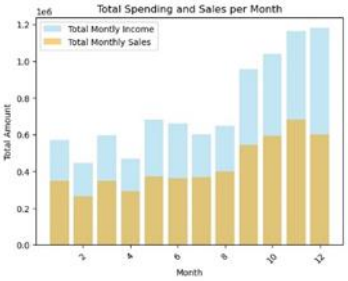


Figure 1

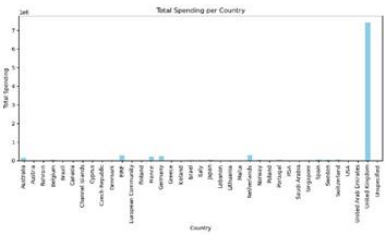


Figure 2

From a strategic perspective, businesses can leverage these insights, focusing on seasonal trends and popular countries can lead the online retail business to manage the inventory in an effective way, optimizing stock levels when demand increases during peak season and planning targeted marketing campaigns.

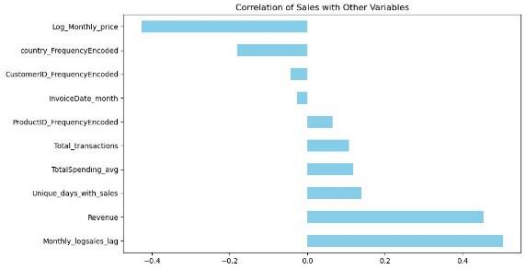


Figure 3

On the other hand, the most strongly correlated variable against sales is the price variable. As can be observed other variables seem to be strongly correlated, but in this case is due to containing the target variables in it, lag variables and revenue are both cases presented. Hence, price is presented as a crucial variable to predict sales.

To further study the effect of prices on sales and linear regression model is built, achieving an R-squared value of 18,2%, indicating that price is an influential but not the only determinant. Other variables should be integrated in the models.

Table IV

| Clusters | P-value | Effect size |
|---|---|---|
| Cluster 0 | 0 | -0.139 |
| Cluster 1 | 1.82E-10 | 0.304 |
| Cluster 2 | 4.90E-181 | -0.193 |
| Cluster 3 | 0 | 0.0323 |

Even though t-statistics values for each cluster indicate significance against the overall market sales, the size effect of clusters suggests that this may not translate into substantial practical implication in this online retail market and their impact on sales strategies could be limited. The cluster distributions are clearly shown in figure 5, where cluster distributions seem to overlay in sales and prices scatterplot. This can also be observed by plotting quadratic regression curves (polynomial models of order 2) for each cluster using the target variable sales and the strongly correlated variable price. It is noticeable that regression curves seem to converge.

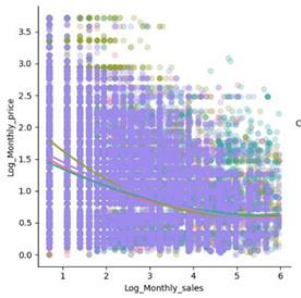


Figure 4

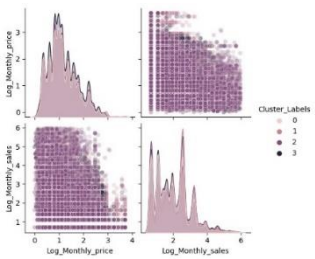


Figure 5

Future analysis incorporating more variables, or the use of different techniques could lead to more pronounced client segmentations. Investigating the response of different clusters to different marketing strategies or personalize customer engagement strategies could open new lines of enquiry.

Table V

| Models | R-squared | RMSE |
|---|---|---|
| XGBoost | 0.692 | 0.545 |
| Linea Regression | 0.328 | 0.805 |

Comparatively, the XGBoost model outperforms the linear regression model, visible in their results. R-squared accounts for 0.69 in XGBoost model and, hence about the 69% of variability in sales can be explain through this model. XGBoost also achieves a moderate prediction error, 0.545, indicating better fit and suggesting that XGBoost model handles better the complexity of the data. However, the linear regression model can only explain the 33% of the variability in sales and achieve a RMSE of 0.805, indicating less accurate predictions. In addition to this, while Linear Regression model focused on Product ID and Customer ID encoded variables, XGBoost uses a broader range of variables focusing on average spending, price, and lag sales variable, leveraging predictions.

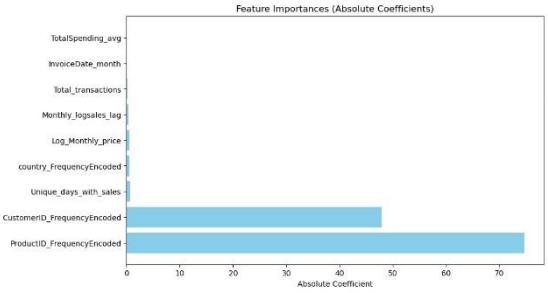


Figure 6

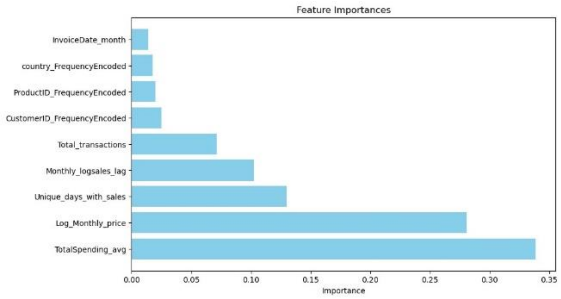


Figure 7

### *A. Limitations*

The study is limited by its temporal scope, which includes only information from 2011. This has a major drawback in that it only records the sales cycle for one year, missing any potential trends or abnormalities from year to year and leaving the model with no future generalization power.

Expanding the dataset and including other relevant variables such as marketing activities, economic factors, and customer demographics could lead to a more comprehensive understanding of sales.
This also could lead the study to build more distinctive client clusters with different purchasing behaviors.

## REFERENCES


[1] M. Ulrich, H. Jahnke, R. Langrock, R. Pesch, and R. Senge, “Classification-based model selection in retail demand forecasting,” International Journal of Forecasting, Jun. 2021.

[2] J. L. Gagnon and J. J. Chu, "Retail in 2010: a world of extremes," Strategy & Leadership, vol. 33, no. 5, pp. 13-23, 2005.

[3] M. C. Cohen, P.-E. Gras, A. Pentecoste, and R. Zhang, Demand Prediction in Retail, vol. 14, ch. 2, "Data Pre-Processing and Modeling Factors," Springer, 2021, p. 22

[4] V. Plevris, G. Solorzano, N. Bakas, and M. Ben Seghier, “Investigation of performance metrics in regression analysis and machine learning based prediction models,” 8th European Congress on Computational Methods in Applied Sciences and Engineering, Jan. 2022.

[5] Praveen K. B., Prateek. J., P. Kumar, and Pragathi. G., "Inventory Management using Machine Learning," Int. J. Eng. Res. Technol., vol. 9, no. 06, Jun. 2020.


| **Section** | **Word count** |
|---|---|
| Abstract | 142 |
| Introduction | 278 |
| Analitical question | 276 |
| Data | 250 |
| Analysis | 987 |
| Results | 589 |
| Total | 2513 |