The Value of Loyalty: Customer impact on Online Store Revenue

Malika Dhakhwa

2024-04-22

# 1. Summary/Abstract

This project investigates the revenue implications of customer return behaviors during the holiday season for a UK-based online retailer specializing in unique gifts. In the current landscape, where e-commerce has reshaped retail, understanding customer loyalty becomes crucial for revenue growth. Utilizing a dataset of transactions spanning from December 2010 to December 2011, this study evaluates the revenue contributions of returning customers versus those who do not return.

The analysis employs Linear, LASSO, and Generalized Additive Models (GAM) to explore the relationship between the sales revenue of the retailer and customer loyalty defined by returns of the customer during holiday season. The models are assessed using the root mean square error (RMSE) metric, with cross-validation reinforcing the analysis. LASSO emerged as the most effective model. The study also addresses the challenges of forecasting customer purchasing patterns using a dataset with limited information.

For more comprehensive understanding of customer behavior, future studies should consider including product-specific information in the analysis and explore the feasibility of accessing demographic information about the customers.

# 2. Introduction

This project examines the revenue dynamics associated with customer returns during the crucial holiday season for a UK-based non-store online retailer specializing in unique all-occasion gifts. Given the transformation of the retail sector driven by online shopping(Khouloud, 2020), which saw global e-commerce sales surpass 5.78 trillion U.S. dollars in 2023 and anticipated to hit 6.3 trillion in 2024(Statista, 2024), understanding customer behaviors becomes imperative for retailers(Chow & Holden, 1997; Heskett et al., 1994). The intersection of advancing IT, increased internet accessibility, and evolving consumer expectations necessitates strategies to foster customer loyalty(Reichheld & Schefter, 2000) as a means to mitigate turnover costs and bolster revenue.

However, the presumed correlation between loyalty and profitability is not always straightforward(Reinartz & Kumar, 2002; Thomas, Reinartz, & Kumar, 2004).While loyal customers are generally seen as valuable due to their repeat business, not every instance of loyalty may translate to noticeable revenue when considering factors like customer service costs and the nature of purchases.Identifying which segment of customers are truly beneficial is crucial for efficient distribution of resources (Reinartz & Kumar, 2002). This project scrutinizes the revenue contributions from customers who returned for holiday shopping compared to those who did not, with the anticipation of reinforcing the notion that loyal customers are inherently good source of revenue generation. The findings aim to offer actionable insights for the retailer to enhance revenue growth strategically.

## 2.1 Background

Building upon the impact of online shopping on consumer behavior and the importance of customer loyalty, we now turn to the subject of study. The unit is a non-store online retail business based on the United Kingdom. It specializes in unique all-occasion gifts and deals with more than three thousand types of merchandises. Its clientele base expands from retail customers to wholesales customers across United Kingdom and other geographical locations.

We have one year of sales data of the unit spanning from December of 2010 to December of 2011. The data reveals a pattern of sales performance with initial months showing steady yet modest results and gaining noticeable momentum as the holiday season approaches. This uptick suggests that the latter part of the year, with its festive attractions, plays a pivotal role in the overall sales activity.

Purchases made outside of major festive events suggests a customer base attracted by the inherent appeal and variety of the product offerings, as opposed to seasonal promotions. In contrast, the peak season not only intensifies competition bringing plethora of special offers but also serves as a test for customer loyalty. It is during this time that returning customers, who are seemingly less swayed by market-wide offers, affirm their status as the cornerstone of the store’s revenue, demonstrating loyalty that extends beyond the allure of holiday discounts. These insights into customer buying patterns are instrumental as we explore the operational subtleties of sales trends and their implications for sustained revenue growth.

## 2.2 Hypothesis

This project is centered around understanding the implication of customer loyalty on revenue generation at a UK-based online retailer specializing in unique gift items. Focusing on a one-year observational period, this study aims to investigate the revenue associated with customer return during the critical holiday season. Our guiding hypothesis is that customers who returned to purchase during this peak shopping period had a disproportionate and positive contribution on the retailer’s revenue streams in comparison to customers who did not make return visits within the same season.

# 3. Methods

## 3.1 Data and Source

The data comprises nine variables capturing essential aspects of transactions such as transaction numbers, product codes, product descriptions, quantities sold, transaction dates, unit prices, customer IDs, the country of residence of the customer, and sales amounts. It encompasses sales data from December 1, 2010, to December 9, 2011. The data does not segregate its sales to wholesalers and retailers.

The dataset was initially acquired from Kaggle for a preliminary assessment of its suitability for this project. However, upon a subsequent visit to the site for additional information, it was discovered that the dataset had been removed from Kaggle.com. A follow-up search on Google revealed the dataset’s availability at the [UC Irvine Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Online+Retail). Notably, the version found in the UC Irvine repository lacks the sales amount column, which is calculable by multiplying the quantities sold by the unit prices. The analysis for this project was conducted using the dataset originally downloaded from Kaggle.

## 3.2 Preliminary Data Cleaning

A key feature of this data is its treatment of each sale of an item on an Invoice as a separate observation, leading to multiple items sold on the same invoice to appear as distinct observations. Initially comprising 541,909 observations, the dataset was subjected to comprehensive cleaning. This process involved removal of records missing crucial customer ID information, conversion of key variables (CustomerID, Country, Description, StockCode) from character to factor types for better manipulation, and the exclusion of observations with negative quantities and sales, indicative of order cancellations.Further scrutiny led to the identification and removal of transactions with unusually high unit prices, likely representing logistical and transportation costs rather than direct transactional values.

Moreover, the dataset’s inability to distinguish sales to wholesalers presented a potential analysis bias due to the inclusion of bulk transactions. To mitigate this, scatter plots were employed to identify and eliminate observations corresponding to abnormally large sales amounts or quantities attributed to a single customer.

Following these cleaning efforts, the dataset was refined to 289,065 observations.For a comprehensive overview of the cleaning process, please refer to the detailed explanation in processing section (R>processing-code.processing\_superstore.qmd).

## 3.3 Feature Engineering

This analysis is comprised of four variables. The variable of interest in this analysis is the cumulative spending of each customer over the transaction period. To assess customer loyalty, which we have identified as our main predictor, we examined whether customers made repeat purchases during the critical holiday season. Additional predictors include the total number of transactions per customer throughout the period and whether the customer resides in the United Kingdom.

We define ‘customer loyalty’ as a dummy variable based on transactional activity. Customers were considered loyal or returning if they had made at least one purchase prior to June 2011 and have returned for at least one additional purchase between October and December 2011. This classification was made based on the invoice date.

To calculate the number of purchases, we aggregated the data by invoice number and subsequently tallied the invoices for each Customer ID. Similarly, the total sum of purchases is derived by aggregating sales figures under each Customer ID.

The final groundwork is consolidated on a per Customer ID basis, ensuring each customer was represented as a unique observation. This step enables a robust statistical analysis of spending patterns, customer loyalty, and geographic influence on purchasing behavior setting the stage for insightful findings on the dynamics of customer engagement and revenue generation.

## 3.4 Model Development

For the predictive analysis, three distinct models were fitted and evaluated: linear models, regularization-based models, and a non-linear model. The dataset was partitioned, allocating 75% for model training purposes and the remaining 25% designated for testing the ultimately selected model. To ensure model robustness, cross-validation was employed across all models, adhering to a ten-fold resampling methodology. The analysis was conducted within the tidymodels framework, with the exception of the cross-validation process for the Generalized Additive Model (GAM), which required handling separately due to its unique characteristics.

## 3.5 Model Overview

### 3.5.1 Linear Model

A linear model was deemed suitable given the limited number of predictor variables and the anticipated linear relationship between these predictors and the outcome variables. This model serves as a benchmark, providing a baseline from which the performance of more complex models can be evaluated.

### 3.5.2 Least Absolute Shrinkage and Selection Operator (LASSO) Regression

In search of a model that balances complexity with interpretability, LASSO regression was selected. LASSO has useful feature of selectiing and effectively reducing the coefficients of less significant predictors to zero. This approach simplifies the model and also aimed at enhancing predictive accuracy.

### 3.5.3 Generalized Additive Model (GAM)

To explore and accommodate potential non-linear relationships between the predictors and outcome variables, the Generalized Additive Model (GAM) was incorporated into the analysis. Chosen for its flexibility, GAM allows for the modeling of complex, non-linear interactions in a way that is both interpretable and visually demonstrable, offering a deeper insight into the underlying data structure.

## 3.6 Model Evaluation Metric

The efficacy of these models was quantitatively assessed using the root mean square error (RMSE). RMSE is the square root of the average squared differences between the predicted and observed values. It offers a comprehensive measure of prediction accuracy, serving as the key metric for model comparison and selection.

# 4. Results

## 4.1 Descriptive analysis

The proportion of customers who purchased in both the periods is 30.72 percent. [Table 1](#tbl-resulttable1) depicts the descriptive statistics of various aspects of customers purchase behavior. The store made 14951 transactions with 4117 customers during the period from December 1, 2010 to December 9, 2011, by recording a mean sales of GBP 290.19 (SD=GBP 187.36) per transaction, an average cumulative sales of GBP 1053.84 (SD = GBP 1429.07) to individual customers and an average sales frequency of above 3 times (SD = 4.28) per customer. The average no. of days before a customer returns to shop is recorded at 97.90 (SD=102.36) days.

Table 1: Descriptive Statistics

| Count | Particulars | Mean | SD | Min | Max |
| --- | --- | --- | --- | --- | --- |
| 14951 | Invoice Value | 290.19 | 187.36 | 0.38 | 1000.00 |
| 4117 | Cumulative Purchase | 1053.84 | 1429.07 | 3.75 | 18282.63 |
| 4117 | Purchase frequency | 3.63 | 4.28 | 1.00 | 45.00 |
| 4117 | Recency of Purchase | 97.90 | 102.36 | 1.00 | 374.00 |

[Figure 1](#fig-result1) presents distribution of sum of customer purchase, invoice value, frequency and recency. For fitting linear model, the sum of purchase and frequency were log transformed due to heavy right-skewness. The metrics were converted into original scale after fitting the model.

|  |
| --- |
| Figure 1: Distribution |

[Figure 2](#fig-result2) depicts the distribution of monthly sales of the store. The sales are noted to be modest with occassional and nominal variations during the initial period. Starting from October, the sales trajectory showed an upward trend, reaching maximum in November, which could be due to early holiday shopping for the approaching Christmas season. It is reasonable to anticipate that past customers would engage in purchasing activities during this peak sales period.

|  |
| --- |
| Figure 2: Distribution of Monthly Sales |

**?@fig-result3** and **?@fig-result4** depict two plots, Recency by Frequency plot and Recency by Monetary plot. While the Recency by Frequency plot revealed an expected pattern where customers with more frequent purchases tended to have more recent transactions, an unexpected trend was also noted. There are some evidence that a segment of customers with a history of multiple purchases did not return during the peak sales period. The Recency by Monetary plot mirrored the trend observed in the ‘Recency by Frequency’ plot. The plot revealed that a subset of customers who have purchased over 2,500 in monetary value in the past, did not make subsequent purchases within the following 100 days.

|  |
| --- |
| Figure 3: Recency Plots |

|  |
| --- |
| Figure 4: Recency Plots |

[Figure 5](#fig-result5) provides a visual impression of the the cumulative purchase by customers who returned during the holiday season (loyal customers) versus those who did not return during the holiday season. It suggests that the returning customers contributed more to the store’s total sales than the non-returning customers.

|  |
| --- |
| Figure 5: Distribution of Cummulative Purchase by Returning status of the Customer |

[Figure 6](#fig-result6) presents country-wise sales trend. Majority of the customers are noted to be from the United Kingdom, followed by Germany, France, Belgium and Spain.

|  |
| --- |
| Figure 6: Countrywise Sales trend |

## 4.2 Basic Statistical Analysis

A simple linear model was fitted to examine the association between total purchase a customer makde and the returning status of the customer. [Table 2](#tbl-resulttable2) depicts the results of the regression. It indicates that there is a positive association between customers returning during the peak season and the cummulative purchase of the customer.

Table 2: Model Metrics

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 5.936122 | 0.0196644 | 301.87197 | 0 |
| loyal\_cust1 | 1.395626 | 0.0352993 | 39.53698 | 0 |

## 4.3 Full Analysis

Three models were fitted to predict the sum of customer purchases: Linear Model, LASSO and GAM. [Table 3](#tbl-resulttable3) shows the model comparison for estimating the sum of purchases. The GAM model has the lowest RMSE, suggesting it performs well in prediction accuracy. However, when subjected to cross-validation, the GAM model’s RMSE revealed a decrease in prediction accuracy across various subsets of the data compared to its initial evaluation based solely on the training data. Both the tuned and untuned Lasso models demonstrated the next lowest RMSE values. We chose the simple Lasso model as our final model for its simplicity.

Table 3: Model comparison

| Model | RMSE |
| --- | --- |
| Simple Linear | 1308.5479 |
| Multivariate Linear | 689.2842 |
| Multivariate Linear Cross-validated | 690.4901 |
| Lasso without Tuning | 675.3834 |
| Lasso - Tuned and Cross-validated | 675.3834 |
| GAM model | 661.0570 |
| GAM Cross-validated | 720.5078 |

The performance of the selected model, the simple Lasso model was evaluated by fitting it on the test set. [Table 4](#tbl-resulttable4) presents the RMSE of the model on the test data.

Table 4: Test Data: Model Metrics

| .metric | .estimator | .estimate |
| --- | --- | --- |
| rmse | standard | 683.8719579 |
| rsq | standard | 0.7522039 |

[Figure 7](#fig-result7) depicts the Predicted versus Observed plot for the training and test data.

|  |
| --- |
| Figure 7: Predicted vs. Observed Plot |

[Figure 8](#fig-result8) presents Residual plot of the test data.

|  |
| --- |
| Figure 8: Residual Plot |

# 5. Discussion

## 5.1 Summary and Interpretation

Three predictive models were applied to compare the total purchase amount by customers who returned during the peak season versus those who did not return: a Linear Model, LASSO, and GAM. Among these, the GAM model exhibited the lowest RMSE, signaling superior prediction accuracy. However, its performance dipped under cross-validation, displaying a decrease in accuracy across various data subsets, which diverged from its initial evaluation based on training data alone.

Both tuned and untuned LASSO models recorded the next lowest RMSE values. This similarity in performance resulted because the tuning process yielded a penalty value for the optimal model close to that of the untuned version, with all predictors being retained. Owing to its simplicity, the simple Lasso model was chosen as the final model. When evaluated on the test set, this model showed a slightly higher RMSE of 683.87, compared to 675.38 on the training set, indicating a consistent yet marginally less accurate prediction capability.

The Predicted versus Observed plots for both training and test data revealed that, for lower-values, the model’s predictions were close to the diagonal line, suggesting relatively accurate predictions. However, as the values increased, both sets of data deviated from the perfect fit line, suggesting diminished accuracy for higher values. This pattern is suggestive of the potential influence of additional factors not captured by the model.

Similarly, the residual plot for the test data revealed a fan-shaped pattern, indicative of heteroscedasticity and variance in prediction errors which increased with higher value predictions. This further points to the existence of other influential factors not considered in the model.

Despite its merits, the Lasso model did not fully capture the complexities of the dataset. The diagnostic plots for both training and test data underscore the need for further model refinement to predict the difference in sum of purchases by returning and not-returning customers during the peak season with more accuracy.

## 5.2 Strengths and Limitations

This project explored a complex relationship between the outcome and predictors by examining Linear, LASSO, and GAM models using limited available information. It also ensured the model’s reliability through cross-validation.

Customer purchasing behavior is inherently multifaceted and deeply influenced by various demographic factors, including but not limited to gender, age, race, and educational background. These aspects can remarkably shape buying habits, preferences, and overall engagement with retail services. In our analysis, the models deployed were constrained by the scope of available data, primarily focusing on purchase frequency, recency, and the total amount spent. This limited information restricts our ability to fully understand and accurately predict customer behavior fully.

Considering the retail store’s extensive inventory of over 3,000 items, along with associated quantities and unit prices, incorporating these information may potentially enhance the model’s accuracy.

## 5.3 Conclusions

Future research could explore how the product-specific data contribute to predicting outcomes more effectively.

# 6. References

Chow, S., & Holden, R. (1997). Toward an understanding of loyalty: The moderating role of trust. *Journal of Managerial Issues*, 275–298.

Heskett, J. L., Jones, T. O., Loveman, G. W., Sasser, W. E., Schlesinger, L. A., et al. (1994). Putting the service-profit chain to work. *Harvard Business Review*, *72*(2), 164–174.

Khouloud, A. (2020). Are we witnessing the awakening of a new world order. *Forbes@ 10*.

Reichheld, F. F., & Schefter, P. (2000). E-loyalty: Your secret weapon on the web. *Harvard Business Review*, *78*(4), 105–113.

Reinartz, W., & Kumar, V. (2002). The mismanagement of customer loyalty. *Harvard Business Review*, *80 7*, 86–94, 125.

Statista. (2024). *Topic: Online shopping*. Statista; <https://www.statista.com/topics/871/online-shopping/#topicOverview>.

Thomas, J. S., Reinartz, W., & Kumar, V. (2004). Getting the most out of all your customers. *Harvard Business Review*, *82*(7-8), 116–123.