Customer Segmentation

Malika Dhakhwa

2024-03-17

# 1. Summary/Abstract

*Write a summary of your project.*

# 2. Introduction

## 2.1 General Background Information

\_Provide enough background on your topic that others can understand the why and how of your analysis

This project aims to explore customer segmentation through the supervised and un-supervised learning perspectives. In the initial phase, the analysis applies and investigates the effectiveness of logistic regression to segregate customers based on the general patterns of customer behavior. As the study advances, it integrates more advanced tools, such as Expectation Maximization algorithms.

It is crucial that organizations understand their customer’s purchasing patterns to design and execute effective and customized engagement strategies. This helps businesses enhance customer service, consequently increase sales. Based on characteristics like spending and purchase frequency, businesses can segment customers and tailor their marketing and sales efforts to meet the unique needs of each group. This personalized approach can enhance customer satisfaction and also improve multiple end results like customer loyalty, operational efficiency, and profitability.

The original data does not have the customers categorized into a group. To use logistic regression, a group variable is created based on the purchase amount and frequency.

## 2.2 Description of data and data source

*Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.*

This study utilizes the ’superstoredata’dataset, focusing on international sales from an online retail store.

Initially this data was downloaded from Kaggle for a preliminary assessment of it’s usability for this project. However, when revisiting the site for further details, it was no longer available on Kaggle.com. A subsequent Google search led to the discovery of the dataset in the [UC Irvine Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Online+Retail). Per the repository, this is a transnational data set spanning from December 1, 2010, to December 9, 2011 from a UK-based, non-store online retail operation. The company specializes in unique all-occasion gifts to a clientele which also includes wholesalers.

This data has nine variables as detailed below:

1. InvoiceNo: It is a categorical variable with a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter ‘c’, it indicates a cancellation.
2. StockCode: It is a categorical variable with a 5-digit integral number uniquely assigned to each distinct product.
3. Description: This represents Product name and is a categorical variable.
4. Quantity: It represents quantities of each product (item) per transaction and is an integer.
5. InvoiceDate: It represents the day and time each transaction was generated.
6. UnitPrice: It represents product price per unit in sterling.
7. CustomerID: It is a categorical variable with a 5-digit integral number uniquely assigned to each customer.
8. Country: It represents the name of the country where each customer resides.
9. Sales: It represents product price multiplied by UnitPrice.

## 2.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

The objective is to understand customer segmentation from logistic regression and advanced analytical tools such as Mixture Models.

# 3. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.*

## 3.1 Schematic of workflow

\_Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

## 3.2 Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

The primary dataset for this project is the ‘superstoredata’ dataset, detailing transactions of an online retailer based in United Kingdom. A key feature of this data is that it counts sales of a particular item from an Invoice as an observation leading to different items sold in the same Invoice to appear as separate observations. Most of the Invoice numbers appear several times in different data observations.

### 3.2.1 Initial Data Cleaning

As the project focuses on international customers, initial data cleaning involved eliminating UK transactions downsizing the data size to 46431 observations across nine variables.

The preliminary inspection identified several issues needing attention such as missing values in ‘CustomerID’, negative ‘signs in ’Quantity’ and ‘Sales’ values, and incorrectly formatted dates. Variables, ‘InvoiceNo’, ‘StockCode’, ‘Description’ and ‘Country’, originally coded as characters, were converted to factor types to facilitate better data manipulation.

1480 observations lacked ‘CustomerID’ and were removed, as customer behavior cannot be tracked without the unique identifier and ‘CustomerID’ was converted to factor type from numeric type. ‘Country’, ‘Description’ and ‘StockCode’ were converted to factors from character variables and transaction date to the date format.

There were 1372 observations with negative values in quantity and sales due to order cancellation which were excluded from the data.

An examination of ‘UnitPrice’ distribution revealed that all the high value goods with a Unitprice above 25 were distributed between 75 and 100 percentile. A closer inspection at this inter-quartile range disclosed that there were 19 product description categories with the Unitprice above 25. Some of those categories had descriptions such as CARRIAGE, POSTAGE, Manual, which appeared to be related to logistics. 1216 observations with such descriptions were dropped from the data since those were less likely to be directly related to customer purchase.

### 3.2.2 Refining the Dataset

The store also sells merchandises in wholesale quantities. A wholesaler potentially had larger purchases and in higher frequencies. In absence of clear information on wholesalers, a scatter plot of Cumulative Invoice Value of customers by Purchase Frequency was plotted to identify potential wholesalers. The scatter plot revealed that most of the customers made purchases less than a frequency of 50 and that most cumulative purchase values were less than 50,000. Assuming that the outliers represented purchases by wholesalers, observations with purchase frequencies above 50 and cumulative Invoice values above 50,000 were dropped from the data. This step excluded 9745 observations from the data. The cleaned data has 32602 observations with 9 variables. Next, the quantities purchased in each invoice were plotted against the individual invoice value. The scatter plot revealed that the observations were concentrated within the purchase quantity below 1000 items and invoice amounts of 2,500. Assuming that the sparsely distributed points represent purchases by wholesalers, observations with Quantity above 1000 and individual Invoice value above 2,500 were further removed from the data.This step excluded 4485 observations for a final total observation of 28117.

The study of customer spending pattern requires the knowledge of each individual transaction amount. The subsequent cleaning process reorganized the dataset by ‘InvoiceNo’, consolidating sales data under each invoice. This final data is saved as processed\_superstore\_RFM in the data > processed-data.

### 3.2.3 RFM Classification

In absence of a labeled categorical variable for classification, the study segmented customers into two distinct groups using a recency, frequency and monetary (RFM) model. RFM model is a commonly used behavior-based model in analysing customer behavior (Yeh, Yang, & Ting, 2009). Recency is measured in days or months and represent the interval between the most recent transaction time and the time of analysis, and a lower number interval is preferred. Frequency is the number of purchases made in a certain period, and monetary is the total amount the customer spent during that time period.(Wei, Lin, & Wu, 2010)

In the RFM classification process, the Recency, Frequency and Monetary values were calculated for each customer, ensuring that each customer was represented only once in the modified dataset.Frequency in this data is the total no. of Invoices issued to the customer, and Monetary is the total purchases made by the customer during the stipulated period. Recency was calculated from the next day of the transaction period. Each customer was assigned a value ranging from 1 to 4 for the Recency, Frequency and Monetary attributes, reflecting their quartile position within each metric. Subsequently, these metrics were combined to formulate a three-digit score for every custoemr, positioning the Monetary value in the hundreds place, Frequency in the tens place, and Recency in the units place. All the customers with a score above 344 were assigned as High-Value Customers.

Following RFM segmentation, logistic regression was applied to predict customer group membership. This method utilized the segments derived from RFM as the dependent variable, with the Recency, Frequency and Monetary attritubtes serving as independent variables. The objective is to develop a model for predictive analysis based on those segments.

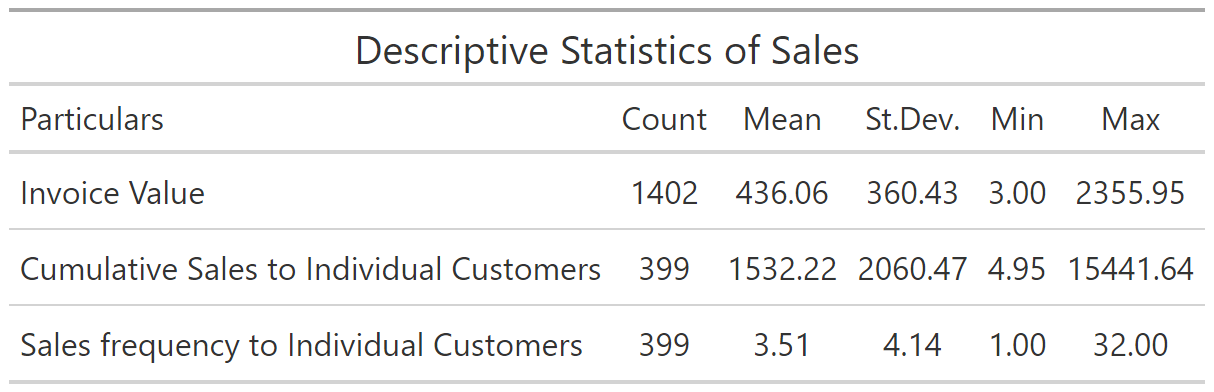
## 3.3 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

**?@tbl-resulttable1** depicts the descriptive statistics of various aspects of customers purchase behavior. There were 1402 transactions with 399 customers outside UK during the period from December 1, 2010 to December 9, 2011. The store recorded a mean sales of GBP 436 per transaction, an average cumulative sales of GBP 1532 to individual customers and sales frequency of above 3 times per customer.



**?(caption)**

[Figure 1](#fig-result1) depicts the distribution of customer segmentation and suggests that approximately 25 percentage of the customers are grouped as Valued Customers.

|  |
| --- |
| Figure 1: Distribution of Sales and Purchase frequency |

[Figure 2](#fig-result2) depicts two density plots for the Cumulative Sales and Invoice Values across customers and a histogram for Customer purchase frequencies. Both density plots are heavily right-skewed.T0he averages of total sales to a customer and total invoice values appear to be less than 2500 and 500 sterling pounds, respectively, supporting the mean values of 1532 and 436 in the descriptive table. The average frequency appears to be less than 5 times, which is also observed in the descriptive statistics table.

|  |
| --- |
| Figure 2: Distribution of Sales and Purchase frequency |

[Figure 3](#fig-result3) depicts a density plot of number of days since the last transaction made by a customer. The distribution is right-skewed. This plot indicates that most customers made a purchase transaction within the last 30 days of the reference date.It also reveals that few customers had not make a transaction for a long time.

|  |
| --- |
| Figure 3: Distribution of Recency of Purchase |

[Figure 4](#fig-result4) depicts a trend of monthly international sales. A spike is noticed in November, presumably due to Christmas in the following month, which is also supported by distribution of the Recency plot.

|  |
| --- |
| Figure 4: Monthly Sales trend |

[Figure 5](#fig-result5) depicts country-wise trend of international sales by sales amount. Top customers were in Germany, France, Spain and Belgium.

|  |
| --- |
| Figure 5: Countrywise Sales trend |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

Penalized logistic regression of customer segment was performed on Recency, Frequency and Monetary variables using ‘tidymodels’ package and 4-fold corss-validation techniques.

[Figure 6](#fig-result6) depicts the area under ROC curve by the range of penalty values. It provides visualization of validation set metrics. The curve has extreme high values, all above 0.9995. A sharp drop is noticed before the penalty value of 0.0010, suggesting the sensitivity of the performance of the model at this penalty value. This is followed by a sharp rise after the penalty value of 0.01. These features of model performance are indicative of model overfitting to the training data and the data being easily separable.

|  |
| --- |
| Figure 6: Visualization of validation set metrics |

[Figure 7](#fig-result7) depicts validation set ROC curve plotting true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The curve moves along 1-specificity and rises only from the end.This suggests that the model has no ability to distinguish between the two customer categories. This suggests that the logistic regression model is not appropriate to the dataset. It pointed out the need to revisit the choice of model.

|  |
| --- |
| Figure 7: Visualization of Validation set ROC curve |

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

#Example **?@tbl-resulttable2** shows a summary of a linear model fit.

**?(caption)**

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 6. References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>

Wei, J.-T., Lin, S.-Y., & Wu, H.-H. (2010). A review of the application of RFM model. *African Journal of Business Management*, *4*(19), 4199.

Yeh, I.-C., Yang, K.-J., & Ting, T.-M. (2009). Knowledge discovery on RFM model using bernoulli sequence. *Expert Systems with Applications*, *36*(3), 5866–5871.