Customer Segmentation

Malika Dhakhwa

2024-04-06

# 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 leverage statistical modeling techniques to predict customer segmentation based on their purchasing behavior pattern, using the Supervised and Machine Learning approaches. In the initial phase, the analysis employs logistic regression to segregate customers based on the general patterns of purchasing behavior. As the study advances, it integrates Machine Learning Tools into the analysis.

It is crucial that organizations understand their customer’s purchasing patterns to devise and implement effective, personalized engagement strategies. By analyzing behaviors such as spending habits and purchase frequency, businesses can segment their customers and tailor marketing and sales efforts to cater to the unique needs of each segment. This personalized approach can not only boost customer satisfaction but also improve multiple end results like customer loyalty, operational efficiency, and profitability.

Given the absence of pre-categorized customer groups in the original data, a grouping variable is engineered based on purchase amount and frequency to facilitate the statistical analysis.

## 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 which is a transactional dataset of a UK-based, non-store online retail operation spanning from December 1, 2010, to December 9, 2011. The store specializes in unique all-occasion gifts to a clientele which includes UK and non-UK based retail and wholesale customers. This study focuses on international sales of the store. The dataset features nine variables, detailing transaction numbers, product codes, product description, quantities sold, transaction dates, unit prices, customer IDs, country of residence of the customer, and sales amounts. Initially this data was downloaded from Kaggle for a preliminary assessment of its 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).

## 2.3 Questions/Hypotheses to be addressed

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

What statistical modeling techniques are most effective in replicating and predicting customer segmentation,based on their purchase behavior pattern, as indicated by purchase amount, purchase frequency and purchase recency?

# 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 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.

### 3.1.1 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 purchase pattern 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.

### 3.1.2 Refining the Dataset

The store also sold 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 at this point had 32602 observations with 9 variables. Next, the quantities purchased in each invoice were plotted against the individual invoice values. 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.

### 3.1.3 RFM Classification

In absence of a pre-existing customer groups, 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 are 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 customer, 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 is applied to predict customer group membership. This method utilized the segments derived from RFM as the dependent variable, with the Recency, Frequency and Monetary attributes serving as independent variables. The objective is to develop a model for predictive analysis based on those segments.

## 3.2 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

[Table 1](#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.

Table 1: Descriptive Statistics

| Particulars | Count | Mean | SD | Min | Max |
| --- | --- | --- | --- | --- | --- |
| Invoice Value | 1402 | 436.06 | 360.43 | 3.00 | 2355.95 |
| Cumulative Sales to Individual Customers | 399 | 1532.22 | 2060.47 | 4.95 | 15441.64 |
| Sales frequency to Individual Customers | 399 | 3.51 | 4.14 | 1.00 | 32.00 |

[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 Valued Customers |

[Figure 2](#fig-result2) depicts a correlation plot for Monetary, Frequency and Recency embedded with density plots. All of the density plots are heavily right-skewed. Monetary amount has a positive correlation of 0.787 with Frequency and a negative correlation of 0.334 with Recency. Likewise, the correlation between Recency and Frequency is negativ with a magnitude of 0.376. The correlation values indicated that collinearity is unlikely to be an issue for the analysis.

|  |
| --- |
| Figure 2: Correlation Plot |

[Figure 3](#fig-result3) 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 3: Monthly Sales trend |

[Figure 4](#fig-result4) depicts country-wise trend of international sales by sales amount. Top customers were in Germany, France, Spain and Belgium.

|  |
| --- |
| Figure 4: 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.*

First, I fitted the traditional logistic regression model to segregate the high value clients without applying the cross-validation. The results returned a warning message that the algorithm did not converge and that the fitted probabilities numerically 0 or 1 occurred. I computed the accuracy and AUC despite this message.

[Table 2](#tbl-resulttable2) depicts the accuracy of and ROC AUC metrics of the model. The model appears to have a perfect accuracy in predicting the customer segment based on the accuracy metric however it does not seem to have not even random guessing ability based on the ROC AUC metric.

Table 2: Model Metrics

| .metric | .estimator | .estimate |
| --- | --- | --- |
| accuracy | binary | 1 |
| roc\_auc | binary | 0 |

## 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.*

### 4.3.1 Penalized Logistic Regression with cross-validation

For further analysis, I fitted a penalized logistic regression to the data using ‘tidymodels’ package and 10-fold cross-validation technique. The goal was to obtain the best value for predictions by tuning the model hyperparametes and using a mixture that chooses a simple model.

The data has 25% of the customers grouped as Valued Customer. Due to this imbalanced proportions of the segment, the data was stratified by the outcome variable for splitting into training and testing sets to ensure that those subsets include similar proportion of Valued Customer.

[Figure 5](#fig-result5) 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 at 1 before the penalty value of 1 followed by a sharp drop, suggesting the sensitivity of the performance of the model at the corresponding penalty value. These features of model performance are indicative of the training data being easily separable.

|  |
| --- |
| Figure 5: Area under ROC curve by Penalty |

[Figure 6](#fig-result6) 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. Contrarily, as in the case of traditional logistic model, the accuracy metrics of this validation set has a high value, 0.9966555.This suggested the need of digging down to the reason of these opposing model performances. The confusion matrix revealed that it almost perfectly identified both classes except for one date point. This indicated a potential inaccuracy or misinterpretation in the ROC curve plotting.

|  |
| --- |
| Figure 6: Visualization of Validation set ROC curve |

I plotted the ROC curve for the predicted probabilities for the negative class i.e. customer category ‘0’ representing ‘Not a Valued Customer’. Using this approach, the ROC-AUC values aligned with the confusion matrix and resulted into a almost perfect curve. The model showed high sensitivity for class 0 and high specificity for class 1, with correctly identifying almost all classes. ### (I need to figure out if this result is valid or interpretable or what caused this. The subsequent parts are assuming this result is valid)

|  |
| --- |
| Figure 7: Validation set using Penalized Logistic Regression |

In the next step, I applied the model to the test data to see its performance on unseen data. The resulting ROC\_AUC curve also suggested that the model performed exceptionally well on the test data in classifying the two classes, with high accuracy, sensitivity and specificity. This excellent performance might potentially due to overfitting, which occurs when a model learns the training data too closely. The data is relatively straightforward. This may have made it easier for the model to learn and predict accurately on the test data.

### 4.3.2 Random Forest with cross-validation

In the next phase, I fitted a Random Forrest model with 10-fold cross validation using a space-filling design to tune, with 25 candidate models and plotted the results of the tuning process. The results indicated that for optimal performance, two or three predictors are required at each node with a varied minimal number of data points for splitting. The model is found to perform exceptionally well with high accuracy.

I compared the validation set ROC curves for the top penalized logistic regression model and random forest model. Both models are found to perform well.

|  |
| --- |
| Figure 8: Validation set: Penalized Logistic Regression vs Fandom Forest |

In the subsequent phase, the random forest model is evaluated on the testing data and is found to perform equally well on the testing data.

# 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.