RFM – CUSTOMER LIFETIME VALUE

Done by

23BM6JP03 - AKHILESH A S

23BM6JP05 - AKSHAY K

23BM6JP14 - ASWIN V T

Introduction

The project's objective is to use online shopping information from a retailer to segment customers according to the RFM framework and subsequently use the RFM metrics to calculate the Customer Lifetime Value using the probabilistic models Pareto/NBD and Gamma-Gamma.

Dataset - Brazilian E-Commerce Public Dataset by Olist

Olist is the largest department store in Brazilian Marketplaces. Olist connects small businesses from across Brazil to multiple channels with ease. Merchants can sell various products to shoppers through the olist website and use logistic partners provided by Olist.

The dataset contains around 100k orders made by online shoppers from 2016 to 2018 at multiple marketplaces in Brazil. The data includes multiple dimensions, including order status, price, payment, customer location, freight information and reviews submitted by customers.

Data Preprocessing

The data existed in different parts related to customer details, geolocation, items, payments, reviews, orders, products, and sellers. They were merged into a single table with just the necessary columns. There were several NULLs in the reviews column. The column was removed since it was unnecessary for our analysis.

RFM - Recency, Frequency and Monetary

The RFM framework is a popular tool in marketing that analyses consumer behavior and subsequently identifies customers to be targeted in various contexts.

- *Recency measures the time that has gone by since a customer last purchased an item from the site.*
- *Frequency measures the number of times a customer has purchased an item.*
- *Monetary measures the total amount of money a customer spends during his whole relationship with the company.*

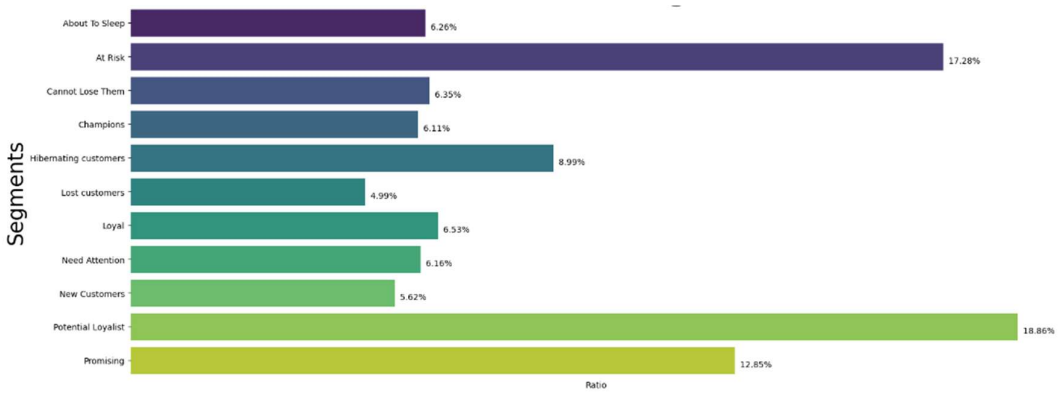
RFM Segmentation

After the RFM metrics were calculated, additional columns were engineered corresponding to each metric, indicating which quintile the customer falls concerning that metric. For example, for a customer with Frequency in the top 20th percentile, the corresponding value would be 1.

According to the Framework, customers are segmented based on the concatenation of the RFM quintiles into 12 different groups. For example, a customer with quintiles 5, 5 and 4 is called a Champion.

Customer Segment	Scores
01: Champions	555, 554, 544, 545, 454, 455, 445
02: Loyal	543, 444, 435, 355, 354, 345, 344, 335
03: Potential Loyalist	553, 551, 552, 541, 542, 533, 532, 531, 452, 451, 442, 441, 431, 453, 433, 432, 423, 353, 352, 351, 342, 341, 333, 323
04: New Customers	512, 511, 422, 421, 412, 411, 311
05: Promising	525, 524, 523, 522, 521, 515, 514, 513, 425, 424, 413, 414, 415, 315, 314, 313
06: Need Attention	535, 534, 443, 434, 343, 334, 325, 324
07: About To Sleep	331, 321, 312, 221, 213, 231, 241, 251
08: Cannot Lose Them But Losing	155, 154, 144, 214, 215, 115, 114, 113
09: At Risk	255, 254, 245, 244, 253, 252, 243, 242, 235, 234, 225, 224, 153, 152, 145, 143, 142, 135, 134, 133, 125, 124
10: Hibernating Customers	332, 322, 233, 232, 223, 222, 132, 123, 122, 212, 211
11: Losing But Engaged	111, 112, 121, 131, 141, 151 **Engagement:** Last email campaign clicked in the last 180 days **OR** Last session_start in the last 90 days
12: Lost Customers	111, 112, 121, 131, 141, 151

Following bar chart indicates the relative size of each segment.



Calculating Customer Lifetime Value

The BG/NBD model is a powerful model for predicting CLV in non-contractual settings (that is, no contract mandates the customer purchase from the vendor). BG/NBD stands for Beta Geometric/Negative Binomial Distribution model. It is an integrated probabilistic model that explains two facets of the behaviours of a consumer, namely the buying behaviour and the churn behaviour.

The following distributions are combined and used to achieve this.

- Exponential distribution is used to model the time between transactions for a single customer.
- The Poisson distribution is used to model the number of transactions completed.
- To account for the variation in buying behaviour across customers, the Gamma distribution is used to model the variation.
- The shifted geometric distribution is used to ascertain whether a customer churns or remains as a customer.
- Like the previous setup, a Beta distribution is used to model the variation in the churn behavior across the populations.

The Gamma-Gamma was used to predict the number of expected purchases made by a customer. According to the BG/NBD model, the value of a customer's transaction follows a Gamma distribution, and the variation of the customer's behaviour across the population also follows a Gamma distribution.

All these assumptions and models are used to determine

- Which customers remain customers and order again in the next period?
- The number of orders these customers would make and their average monetary value.

When implementing the BG/NBD model, Grid Search was used to find the best L2 coefficient using RMSE as the metric.

Combining these models, CLV was calculated for each customer.

K-Means Clustering

Using the RFM metrics and the CLV calculated above, the customers were clustered using the K-Means algorithm. The elbow method was used to identify the ideal number of clusters to be 4. Distortion square, the sum of squares of the distance between each point and its centroid, was used to tune the number of clusters.

The clustering achieved a silhouette score of 0.67.