SUNANDAN DIVATIA SCHOOL OF SCIENCE, NMIMS

Department of Statistics

M.Sc. Statistics & Data Science

Project Report on

CUSTOMER BEHAVIOUR MODELLING

Project Team:

Adarsh Baldawa – 75322100053

Devendra Deshmane – 75322100031

Harsh Junagade – 75322100038

Neeyati Satgounda – 75322100057

Parag Jadhav – 75322100002

Sukhada Sakhalkar – 75322100032

Project Guide:

Dr. Pradnya Khandeparkar

Acknowledgement

Successful accomplishment of this project lies in the hands of everyone who have helped us in this endeavour. We take this opportunity to thank all of them.

We are thankful to the Head of the Department of Statistics, Sunandan Divatia School of Science, NMIMS, Prof. Sunil Shirvaikar for giving us this opportunity to work on this project.

We express our heartfelt gratitude to our guide Dr. Pradnya Khandeparkar for her guidance and support throughout the project. We are extremely thankful to all the Teachers of Department of Statistics, Sunandan Divatia School of Science, NMIMS for their valuable suggestions and encouragement.

Our appreciations also go to our all friends who have willingly helped us out with their abilities. Lastly, we are eternally grateful to our parents and family for their strong backing and promising support.

Index

Introduction

Customer Lifetime Value (CLTV) is used as a collective term to refer to a broad range of economic services provided by the finance industry, which encompasses a broad range of organizations that manage money, including credit unions, banks, credit card companies, insurance companies, consumer finance companies, stock brokerages, investment funds. Customer lifetime value is a prediction of the net profit attributed to the entire future relationship with a customer. The definition clearly states that Customer lifetime value modelling is, calculating how much a customer can bring to the revenue of a company during his/her lifetime. Moreover, it is a calculated figure which is predicted by the customer's purchase and interaction history with the eCommerce website (or any other businesses)

The case we are studying concerns a UK-based and registered online retail store without any brick-and-mortar stores. The decision-makers must target customer groups and develop market strategies to satisfy customer needs and thereby increase the market share of the company. For this study, one year of data on purchase transactions has been collected. Our main goals are to create customer segments using RFM and predicting future CLTV. After segmenting customers, we get an idea to focus on which customers, whom to provide discounts and etc. based on their R-F-M ranks.

Motivation

There are 2 types of transaction, one is contractual (e.g., paying bills, paying taxes etc.) and other transaction is non-contractual (e.g., shopping, booking holidays etc.)

1. There are many companies competing in this domain, and to be relevant, profitable and to grow further companies need to increase their sales.
2. There are many ways to increase the sales of a company. Some of the ways are to:
    - Acquire a new customer
    - Reward the loyal customers.
3. Blindly trying to target customers is harmful for a business, therefore, we should know which customers to target and which we should not target.
4. Calculating retention rates for non-contractual relationships is unfeasible, so we need to devise a method for the marketing team to properly target these customers.

Objectives

- To determine current customer value using RFM method.
- To segment customers into small groups and addressing individual customers based on actual purchasing behaviors.
- To predict future Customer Lifetime Values for 3, 6 and 12 months and creating loyalty segments using those values.

Data Description

Data Source:

This is a transactional data set that contains all the transactions occurring between 01/12/2009 and 09/12/2010 (DD/MM/YYYY) for a UK-based and registered online retail store without any brick-and-mortar stores. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. This data contains 8 attributes. This is secondary data collected from analyticsvidya.com.

	A	B	C	D	E	F	G	H
1	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
2	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	01-12-2009 07:45	6.95	13085	United Kingdom
3	489434	79323P	PINK CHERRY LIGHTS	12	01-12-2009 07:45	6.75	13085	United Kingdom
4	489434	79323W	WHITE CHERRY LIGHTS	12	01-12-2009 07:45	6.75	13085	United Kingdom
5	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	01-12-2009 07:45	2.1	13085	United Kingdom
6	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	01-12-2009 07:45	1.25	13085	United Kingdom
7	489434	22064	PINK DOUGHNUT TRINKET POT	24	01-12-2009 07:45	1.65	13085	United Kingdom
8	489434	21871	SAVE THE PLANET MUG	24	01-12-2009 07:45	1.25	13085	United Kingdom
9	489434	21523	FANCY FONT HOME SWEET HOME DOORMAT	10	01-12-2009 07:45	5.95	13085	United Kingdom
10	489435	22350	CAT BOWL	12	01-12-2009 07:46	2.55	13085	United Kingdom
11	489435	22349	DOG BOWL , CHASING BALL DESIGN	12	01-12-2009 07:46	3.75	13085	United Kingdom
12	489435	22195	HEART MEASURING SPOONS LARGE	24	01-12-2009 07:46	1.65	13085	United Kingdom
13	489435	22353	LUNCHBOX WITH CUTLERY FAIRY CAKES	12	01-12-2009 07:46	2.55	13085	United Kingdom
14	489436	48173C	DOOR MAT BLACK FLOCK	10	01-12-2009 09:06	5.95	13078	United Kingdom
15	489436	21755	LOVE BUILDING BLOCK WORD	18	01-12-2009 09:06	5.45	13078	United Kingdom
16	489436	21754	HOME BUILDING BLOCK WORD	3	01-12-2009 09:06	5.95	13078	United Kingdom
17	489436	84879	ASSORTED COLOUR BIRD ORNAMENT	16	01-12-2009 09:06	1.69	13078	United Kingdom
18	489436	22119	PEACE WOODEN BLOCK LETTERS	3	01-12-2009 09:06	6.95	13078	United Kingdom
19	489436	22142	CHRISTMAS CRAFT WHITE FAIRY	12	01-12-2009 09:06	1.45	13078	United Kingdom
20	489436	22296	HEART IVORY TRELLIS LARGE	12	01-12-2009 09:06	1.65	13078	United Kingdom

Description of Data Attributes:

1. **InvoiceNo:** This is invoice number for the products purchased by customer. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter ‘c’, it indicates a cancellation.
2. **StockCode:** This is product code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
3. **Description:** This is product name. Nominal.
4. **Quantity:** This is the quantities of each product per transaction. Numeric.
5. **InvoiceDate:** This is invoice date and time. Numeric, the day and time when each transaction was generated.

6. **Price:** This is unit price of products. Numeric, Product price per unit is in pounds.
7. **CustomerID:** This is customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
8. **Country:** Nominal, the name of the country where each customer resides.

Data Pre-processing and Cleaning:

- There is no "Total Price" column in the dataset, so we added "Total Price" column by multiplying quantity and price columns.

	A	B	C	D	E	F	G	H
1	Invoice	StockCode	Quantity	InvoiceDate	Price	Customer ID	Country	Total Price
2	489434	85048	12	01-12-2009 07:45	6.95	13085	United Kingdom	83.4
3	489434	79323P	12	01-12-2009 07:45	6.75	13085	United Kingdom	81
4	489434	79323W	12	01-12-2009 07:45	6.75	13085	United Kingdom	81
5	489434	22041	48	01-12-2009 07:45	2.1	13085	United Kingdom	100.8
6	489434	21232	24	01-12-2009 07:45	1.25	13085	United Kingdom	30
7	489434	22064	24	01-12-2009 07:45	1.65	13085	United Kingdom	39.6
8	489434	21871	24	01-12-2009 07:45	1.25	13085	United Kingdom	30
9	489434	21523	10	01-12-2009 07:45	5.95	13085	United Kingdom	59.5
10	489435	22350	12	01-12-2009 07:46	2.55	13085	United Kingdom	30.6
11	489435	22349	12	01-12-2009 07:46	3.75	13085	United Kingdom	45
12	489435	22195	24	01-12-2009 07:46	1.65	13085	United Kingdom	39.6
13	489435	22353	12	01-12-2009 07:46	2.55	13085	United Kingdom	30.6
14	489436	48173C	10	01-12-2009 09:06	5.95	13078	United Kingdom	59.5
15	489436	21755	18	01-12-2009 09:06	5.45	13078	United Kingdom	98.1
16	489436	21754	3	01-12-2009 09:06	5.95	13078	United Kingdom	17.85
17	489436	84879	16	01-12-2009 09:06	1.69	13078	United Kingdom	27.04
18	489436	22119	3	01-12-2009 09:06	6.95	13078	United Kingdom	20.85
19	489436	22142	12	01-12-2009 09:06	1.45	13078	United Kingdom	17.4
20	489436	22296	12	01-12-2009 09:06	1.65	13078	United Kingdom	19.8

- We checked the missing values in the data and found missing values in 'Description' and 'CustomerID' columns. Since our objective is to identify customer groups, there should not be any missing values in 'CustomerID' column. Hence, we removed the NA values since the data without 'CustomerID', since they were of no use for us.

	count	mean	std	min	25%	50%	75%	max
Quantity	525461.0	10.337667	107.424110	-9600.00	1.00	3.0	10.00	19152.00
Price	525461.0	4.688834	146.126914	-53594.36	1.25	2.1	4.21	25111.09

- We found out that there were negative values in the quantity column. We filtered all the values of the dataset and observed that there were negative values. Since we are trying to categorize our customers, we removed all the negative values as these values represent the refunds/returns by the customers.

Methodology

RFM Model:

RFM (recency, frequency and monetary) model is a behaviour-based model used to analyse the behaviour of a customer and then make predictions based on the behaviour in the database. Furthermore, these three variables (R, F and M) belong to the behavioural variables and can be used as the segmenting variables by observing customers' attitudes toward the product, brand, benefit, or even loyalty from the database.

Definitions of Recency, Frequency and Monetary:

1. **Recency(R):** Recency represents the interval between the latest buying date of a customer and the last date of the dataset. The lower the interval the higher is the value of R of the particular customer.

   A customer that has recently shopped at a website is more likely to shop again than a customer who hasn't made a purchase in quite some time. This data can be used to target users that have not shopped in a while by offering them a special discount with an email/message encouraging them to give a try to the website. If a user has recently shopped, we can offer them a retargeting ad with an item that might complement whatever they recently purchased.
2. **Frequency(F):** Frequency represents the number of times a customer buys within a particular period given in the dataset like once in a month, thrice in a year. The higher the value number of transactions in an interval the higher is the value of F of the particular customer.

   How often a customer buys something on the website can be affected by many factors like price point, the need for replacement, and obviously, the type of product. We can use the various digital strategies to direct users to replace their product based on a predicted purchase cycle.
3. **Monetary(M):** Monetary is the total value of transacted sales by each customer, i.e., how much money a customer has spent on purchases. It is the profitability value of the customer's relationship with the company. Higher the value of transacted sales, more is the value of M.

   Monetary value has to do with how much a customer has spent during a particular length of time. Just like in Vegas, "high rollers" usually get special treatment as the ROI for

customers with a high monetary value tend to be greater. We can't disregard customers with low monetary value as they may be the best cheerleaders as loyal customers.

To conduct RFM analysis, we need to rank the customers based on each RFM attribute separately. For customer segmentation, we rank those customers from 1-5 using RFM values. We use the scoring model for ranking those RFM values.

Scoring Model:

The RFM model is the most frequently adopted segmentation technique that comprises three measures, i.e., recency, frequency and monetary. Each of these measures is divided in five equal quintiles each containing 20% of customers. This is known as quintile discretization method. The R, F and M ranks are then combined to form a three-digit RFM cell code.

Among the three RFM measures, recency is often regarded as the most important one. However, according to prior findings, RFM values are inclined to be firm-specific and are based on the nature of the products. For example, Fader et al. (2005) found that for lower recency, customers with higher frequency tended to have lower future purchasing potential than those with lower pre-purchasing rates. Lumsden et al. (2008) have similar findings that there are significant differences between groups across recency and frequency.

The process to quantify customer behaviour via RFM model is as follows:

1. First, sort the database by each dimension of RFM and then divide the customer list into five equal segments. Different RFM quintiles have different response rates.
2. For recency, customers are sorted by purchase dates. Recency is commonly defined by the number of periods since the last purchase, which measures the interval between the most recent transaction date and the final date of the dataset (in days), that is, the lower the number of days, the higher the score of recency. A customer having a high score of recency implies that he or she is more likely to make a repeat purchase. The top 20% segment is coded as 5, while the next 20% segment is coded as 4 and so on. Finally, the recency for each customer in the database is denoted by a rank from 5 to 1.
3. For frequency, the database is sorted by purchase frequency (the number of purchases) made in a certain time period. The definition of frequency is often simplified to consider two states, including single and repeated purchases. The top quintile is assigned a value of 5 and the others are given the values of 4, 3, 2 and 1. Higher frequency score indicates

greater customer loyalty. A customer having a high score of frequency implies that he or she has great demand for the product and is more likely to purchase the products repeatedly.

4. The score of frequency is defined with regards to the fact that the single purchasers are assigned a score of 1. Then, the average of the remaining frequency values is used to determine the mean. Once a customer's total frequency value is lower than the mean, a score of 2 is given to this customer. The process may be repeated more than two times.
5. For monetary, customers are coded by the total amount of money spent during a specified period of time. The definition of monetary is defined by the dollar value that the customer has spent in this time period or by the average dollar amount per purchase or all purchases to date. Marcus (1998) suggested that it is better to use the average purchase amount rather than the total accumulated purchase amount so as to reduce co-linearity of frequency and monetary. Five quintiles are still created and each has equal amounts of sales.
6. Finally, all customers are presented by 555, 554, 553, …, 111, which thus creates 125 (5×5×5) RFM cells. Moreover, the best customer segment is 555, whereas the worst customer segment is 111. Based on the assigned RFM behaviour scores, customers can be grouped into segments and their profitability can be further analysed.
7. In addition to use the value of each cell to judge whether the customer is valuable, four segments are created by using composite scores. The composite scores of RFM is obtained by adding normalized RFM scores of each customer and the weight of RFM variables. Each measure of RFM has the same weight when calculating a composite score. For instance, for the cell (5, 2, 4), the composite score is (5+2+4=) 11.

Customer Segmentation using RFM values:

This step divides customers into numerous groups with similar RFM values, and assigns each customer to an appropriate segment. RFM analysis is used to evaluate customer loyalty, and thus identify the target customers with high RFM values. The main advantage of this process is to be able to adopt different marketing strategies for different customer segments.

First, we found RFM values and then based on these values we ranked these customers from 1-5, 1 being the lowest and 5 being the highest. We get 1-5 ranks for each attribute of R F and M. Using R and F ranks, we divided our customers into 10 small segments (according to industry standards) named 'Hibernating', 'At Risk', 'Can't Lose', 'About to Sleep', 'Need Attention', 'Loyal Customers', 'Promising', 'New Customers', 'Potential Loyalists', 'Champions'.

Secondly, we used R F and M ranks to get a composite score with a minimum value of 3 (R=1, F=1, M=1) and maximum value of 15 (R=5, F=5, M=5) which are then divided into 4 bins of composite scores of equal lengths which we call as loyalty levels, i.e., Platinum, Gold, Silver and Bronze. Given below are the characteristics of each loyalty level.

i. Platinum (Composite score(s): 13,14,15): People in this group are more frequent buyers. These are your most loyal customers, who bought recently, most often, and are heavy spenders. Reward these customers so that they can become an early adopter for your future products and help to promote your brand.
ii. Gold (Composite score(s): 10,11,12): People in this group are average buyers. These are your recent customers with an average frequency and who spent a good amount. Offer membership or loyalty programs or recommend related products to upsell them and help them become your A-tier members.
iii. Silver (Composite score(s): 7,8,9): People in this group are buyers who bought the least. These are your customers who purchased a decent number of times and spent good amounts but haven't purchased recently. Sending them personalized campaigns, offers, and product recommendations will help to reconnect with them.
iv. Bronze (Composite score(s): 3,4,5,6): This is the dormant group. These are customers who used to visit and purchase in your platform but haven't been visiting recently. Bring them back with relevant promotions, and run surveys to find out what went wrong and avoid losing them to a competitor.

Predicting future CLTV:

CLTV is a measurement of how valuable a customer is to your company, not just on a purchase-by-purchase basis but across the whole relationship. Probabilistic lifetime value estimation is made with time projection for a certain 't' time. CLTV is a dynamic concept, not a static model. The most basic formula we use is as follows:

CLTV = Expected Number of Transactions x Expected Average Profit

We will estimate the "Expected Number of Transaction" part using the BG/NBD model and the "Expected Average Profit" part using the gamma-gamma sub-model.

BG/NBD Model:

Beta Geometric / Negative Binomial Distribution is also known as BG-NBD Model. It belongs to the "Buy Till You Die" family of models. It gives us the expected number of transactions in a future period of length t.

This model, models 2 processes by using probability for predicting the expected number of transactions:

- Transaction Process (Buy)
- Dropout Process (Till You Die)

BG/NBD Assumptions

The BG/NBD model is based on the following five assumptions:

1. While active, the number of transactions made by a customer follows a Poisson process with transaction rate λ. This is equivalent to assuming that the time between transactions is distributed exponentially with transaction rate λ, i.e.,
   $f(t_j|t_{j-1};\ \lambda) = \lambda e^{-\lambda(t_j - t_{j-1})};\ \ t_j > t_{j-1} \geq 0$

2. Heterogeneity in λ follows a gamma distribution with pdf
   $f(\lambda|r,\alpha) = \frac{a^r \lambda^{r-1} e^{-\lambda a}}{\Gamma(r)},\ \lambda > 0$

3. After any transaction, a customer becomes inactive with probability p. Therefore, the point at which the customer "drops out" is distributed across transactions according to a (shifted) geometric distribution with probability mass function.
   $P(\text{inactive immediatedly after jth transaction}) = p(1-p)^{j-1},\ j = 1,2,3,\ldots$

4. Heterogeneity in p follows a beta distribution with
   $f(p|a,b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)},\ 0 \leq p \leq 1\ \ (2)$
   where $B(a,b)$ is the beta function, which can be expressed in terms of a gamma function:
   $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$

5. The transaction rate λ and the dropout probability p vary independently across customers.

Model Development at the Individual Level

Likelihood Function to estimate parameters:

$$L(\lambda, p \mid X = x, T) = (1-p)^x \lambda^x e^{-\lambda T} + \delta_{x>0} p(1-p)^{x-1} \lambda^x e^{-\lambda t_x}$$

where $\delta_{x>0} = 1$ if $x > 0$.

Derivation of $P(X(t) = x)$:

$$P(X(t) = x \mid \lambda, p) = (1-p)^x \frac{(\lambda t)^x e^{-\lambda t}}{x!} + \delta_{x>0} p(1-p)^{x-1} . \left[1 - e^{-\lambda t} \sum_{j=0}^{x-1} \frac{(\lambda t)^j}{j!}\right]$$

Derivation of $E[X(t)]$:

$$E[X(t) \mid \lambda, p] = \lambda t \cdot p(\tau > t) + \int_0^t \lambda \tau g(\tau \mid \lambda, p)\, d\tau = \frac{1}{p} - \frac{1}{p} e^{-\lambda p t}$$

Model Development for a Randomly Chosen Individual:

$$E(Y(t) \mid X = x, t_x, T, r, \alpha, a, b) = \frac{\frac{a+b+x-1}{a-1}\left[1 - \left(\frac{\alpha+T}{\alpha+T+t}\right)^{r+x} \cdot {}_2F_1\left(r+x, b+x; a+b+x-1; \frac{t}{\alpha+T+t}\right)\right]}{1 + \delta_{x>0} \frac{a}{b+x-1} \left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x}}$$

Where ${}_2F_1$ is the Gaussian hypergeometric function:

$${}_2F_1\left(r+x, b+x; a+b+x-1; \frac{t}{\alpha+T+t}\right) = \frac{\Gamma(a+b+x-1)}{\Gamma(r+x)\Gamma(b+x)} \sum_{j=0}^{\infty} \frac{\Gamma(r+x+j)\Gamma(b+x+j)}{\Gamma(a+b+x-1+j)} \times \frac{\left(\frac{t}{\alpha+T+t}\right)^j}{j!}$$

Parameters:

- x refers to frequency for each customer.
- t_x refers to time of last purchase for any given customer.
- T refers to the time span of the data set (days).
- r, α comes from the gamma distribution. **Transaction rate of the mass.**
- a, b comes from the beta distribution. **The dropout rate of the mass.**
- X(t) refers to the expected number of transactions in a future period of length t.
- Y(t) refers to the expected number of transactions **for each customer** in a future period of length t.

Note: We can proceed with gamma-gamma sub-model if and only if the correlation between monetary and frequency is negligible.

Gamma-Gamma sub-model:

We use this model for predicting how much average profit we can earn for each customer. It gives us the expected average profit for each customer after modelling the average profit for the mass.

Our model of spend per transaction is based on the following three general assumptions:

- The monetary value (e.g., $, £) of a customer's given transaction varies randomly around their average transaction value.
- Average transaction values vary across customers but do not vary over time for any given individual.
- The distribution of average transaction values across customers is independent of the transaction process.

For a customer with x transactions, let $z_1, z_2, \ldots, z_x$ denote the value of each transaction. The customer's observed average transaction value by

$$\bar{z} = \left(\sum_{i=1}^{x} z_i \right) / x$$

$\bar{z}$ is an imperfect estimate of their (unobserved) mean transaction value ζ. Our goal is to make inferences about ζ given $\bar{z}$, which we denote as $E(Z|\, \bar{z}, x)$. As a first step, we need to derive the distribution of $\bar{z}$ given x transactions.

Model Development

Assumptions:

1. $z_i \sim$ gamma(p, ν), with $E(Z_i \mid p, \nu) = \zeta = p/\nu$.
    - given the convolution properties of the gamma, it follows that total spend across x transactions is distributed gamma(px, ν).
    - given the scaling property of the gamma distribution, it follows that $\bar{z} \sim$ gamma(px, νx).
2. $\nu \sim$ gamma(q, γ).

Deriving f($\bar{z}$|x):

Given these assumptions, the distribution of $\bar{z}$ given x is

$$f(\bar{z}|p,q,\gamma;x) = \frac{1}{\bar{z}B(px,q)}\left(\frac{\gamma}{\gamma+x\bar{z}}\right)^{q}\left(\frac{x\bar{z}}{\gamma+x\bar{z}}\right)^{px}$$

Parameter Estimation:

Given the frequency (x_i) and monetary value ($\bar{z}_i$) data for each individual (i = 1, . . ., I), the sample log-likelihood function is simply,

$$LL(p,q,\gamma|\text{ data}) = \sum_{i=1}^{I} \ln\left[f(\bar{z}_i|p,q,\gamma;x_i)\right]$$

Deriving f(ζ):

We denote a customer's (unobserved) mean transaction value by ζ. conditional on p and ν, ζ = p/ν. However, since ν varies across customers according to a gamma(q, γ) distribution, we view the (unobserved) mean transaction value as a random variable Z with realization ζ.

$$f(\zeta|p,q,\gamma) = \frac{(p\gamma)^{q}\zeta^{-q-1}e^{-\frac{p\gamma}{\zeta}}}{\Gamma(q)}$$

Deriving E(Z| $\bar{z}$, x):

$$E(Z|\,p,q,\gamma;\bar{z},x) = \left(\frac{q-1}{px+q-1}\right)\cdot\frac{p\gamma}{q-1} + \left(\frac{px}{px+q-1}\right)\bar{z}$$

- x refers to frequency for each customer.
- $\bar{z}$ refers to the monetary for each customer.
- Z refers to the expected value of transactions (expected average profit)
- p, q, γ comes from the gamma distribution.

CLTV = Expected Number of Transaction * Expected Average Profit

$$CLTV = E(Y(t)|\,X = x, t_x, T, r, \alpha, a, b) \times E(Z|\,p,q,\gamma;\bar{z},x)$$

$$CLTV = \left[\frac{\frac{a+b+x-1}{a-1}\left[1-\left(\frac{\alpha+T}{\alpha+T+t}\right)^{r+x}\cdot{}_2F_1\left(r+x,b+x;a+b+x-1;\frac{t}{\alpha+T+t}\right)\right]}{1+\delta_{x>0}\frac{a}{b+x-1}\left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x}}\right] \times \left[\left(\frac{q-1}{px+q-1}\right)\cdot\frac{p\gamma}{q-1} + \left(\frac{px}{px+q-1}\right)\bar{z}\right]$$

Segmentation of predicted CLTV:

The method that has been used for segmenting future CLTV values is identical to the one that we used in segmentation of RFM composite scores. The segmentations are not very informative but the descriptive statistics of the loyalty levels give us a basic insight on the attributes/characteristics/habits of each loyalty level.

Results

RFM Model:

Output Table (Showing for 20 customers out of 4312**):**

Customer ID	Recency	Frequency	Monetary	Recency Score	Frequency Score	Monetary Score	RFM Combination	Segment	RFM Score	Loyalty Level
12346	165	11	372.86	2	5	2	252	Can't Loose	9	Silver
12347	3	2	1323.32	5	2	4	524	Potential Loyalists	11	Gold
12348	74	1	222.16	2	1	1	211	Hibernating	4	Bronze
12349	43	3	2295.02	3	3	5	335	Need Attention	11	Gold
12351	11	1	300.93	5	1	2	512	New Customers	8	Silver
12352	11	2	343.8	5	2	2	522	Potential Loyalists	9	Silver
12353	44	1	317.76	3	1	2	312	About to Sleep	6	Bronze
12355	203	1	488.21	1	1	2	112	Hibernating	4	Bronze
12356	16	3	3562.25	4	3	5	435	Potential Loyalists	12	Gold
12357	24	2	11266.93	4	2	5	425	Potential Loyalists	11	Gold
12358	11	3	2703.71	5	3	5	535	Potential Loyalists	13	Platinum
12359	61	6	2537.34	3	4	5	345	Loyal Customers	12	Gold
12360	15	5	1569.41	5	4	4	544	Champions	13	Platinum
12361	98	3	321.35	2	3	2	232	At Risk	7	Silver
12362	374	1	36.94	1	1	1	111	Hibernating	3	Bronze
12366	269	1	500.24	1	1	3	113	Hibernating	5	Bronze
12368	264	1	917.7	1	1	3	113	Hibernating	5	Bronze
12369	49	3	1769.73	3	3	4	334	Need Attention	10	Gold
12370	260	3	756.26	1	3	3	133	At Risk	7	Silver

Using definitions of R, F and M we assigned R, F and M scores to each customer. A combination of the three scores was segmented into ten categories (industry standard):

Hibernating	1015
Loyal Customers	742
Champions	663
At Risk	611
Potential Loyalists	517
About to Sleep	343
Need Attention	207
Promising	87
Can't Loose	77
New Customers	50

Then the sum of R, F and M scores was taken to get the composite scores which were segmented into 4 (arbitrary) parts:

Platinum	1078
Gold	1078
Silver	1078
Bronze	1078

Loyalty level	Recency	Frequency	Monetary	RFM score
Bronze	188.626806	1.152488	258.840547	4.662119
Silver	84.74627	2.079535	662.466114	7.955317
Gold	47.592955	4.078278	1571.780261	10.953033
Platinum	15.823748	12.196757	6219.891555	14.064865

Platinum: People in this group are more frequent buyers with average days since the last purchase is 16 and the average number of times they have transacted in the platform is about 12 times in the last 1 year. Also, their average sales value is 6,220 pounds.

Gold: This group has an average frequency of 4 times and average recency of 47 days. This group is also high spenders with average sales of about 1,572 pounds.

Silver: People in this group have made a transaction on average about 85 days ago. Their average frequency and average monetary values are 2 times and 662 pounds respectively.

Bronze: This is the dormant group with average days since their last purchase is 188. They have transacted on an average 1 time with average sales of 258 pounds.

BG/NBD Model:

T = 374 days

Parameters Estimation:

The model was fitted with 2893 subjects, estimated values of the parameters are:

a: 1.68, α: 0.34, b: 6.68, r: 0.33

Output Table (Showing for 20 customers out of 2893)**:**

Customer ID	t_x	P(Alive)	E[Y(3 months)]	E[Y(6 months)]	E[Y(12 months)]
12346	165	0.00141309	0	0.01	0.01
12347	3	0.013627595	0.05	0.07	0.1
12349	43	0.020964428	0.02	0.04	0.08
12352	11	0.351457716	1.48	2.14	2.82
12356	16	0.062649588	0.22	0.33	0.47
12357	24	0.820526291	3.77	5.33	6.9
12358	11	4.85705E-05	0	0	0
12359	61	7.96186E-05	0	0	0
12360	15	9.79229E-07	0	0	0
12361	98	0.093913799	0.08	0.15	0.27
12369	49	0.016251391	0.02	0.03	0.05
12370	260	0.754493283	0.68	1.26	2.21
12371	45	0.011748487	0.01	0.02	0.04
12373	260	0.753708529	0.47	0.89	1.59
12374	57	0.052284263	0.06	0.11	0.19
12376	25	0.446997519	1.24	1.95	2.81
12377	17	0.011540867	0.01	0.02	0.03
12379	58	0.274011638	0.3	0.53	0.9
12380	101	0.002757136	0.01	0.01	0.02

We found out the probability of being alive/active at any given time t. We also calculated expected number of transactions for each customer at time t=3, 6 and 12 months.

Correlation Matrix:

	frequency	monetary
frequency	1.000000	0.138881
monetary	0.138881	1.000000

We can observe that the correlation values between frequency and monetary is 0.138881. Any correlation lower than 0.2 is considered negligible. Hence, we can go forward with the gamma-gamma sub-model.

Gamma-Gamma Sub-Model:

T= 374 days

Parameters Estimation:

The model was fitted with 2893 subjects, estimated values of the parameters are:

p: 12.23, q: 0.93, v: 12.14

Output Table (Showing for 20 customers out of 2893)**:**

Customer ID	CLTV(3 months)	CLTV(6 months)	CLTV(12 months)	segment_3_month	segment_6_month	segment_12_month
12932	35.56648446	62.68506812	101.2974178	Gold	Gold	Gold
12933	0.000105315	0.000191303	0.000323199	Silver	Silver	Silver
12935	4.46135E-08	7.69909E-08	1.20918E-07	Bronze	Bronze	Bronze
12937	2.055382905	3.738441893	6.328277708	Gold	Gold	Gold
12939	1.14189E-12	2.0027E-12	3.21244E-12	Bronze	Bronze	Bronze
12943	145.1420628	263.7009695	446.6433561	Platinum	Platinum	Platinum
12948	55.85185838	104.0682691	183.1595668	Gold	Gold	Gold
12949	5.22939E-15	9.22671E-15	1.49216E-14	Bronze	Bronze	Bronze
12951	4.98822E-08	8.83243E-08	1.4363E-07	Bronze	Bronze	Bronze
12956	83.56518164	153.9041228	265.7907521	Platinum	Platinum	Platinum
12957	133.6354302	242.7015723	409.9173609	Platinum	Platinum	Platinum
12958	15.22066471	27.47111038	46.08527397	Gold	Gold	Gold
12960	0.937066521	1.689374357	2.823800766	Silver	Silver	Silver
12963	5.47906E-08	9.20714E-08	1.39998E-07	Bronze	Bronze	Bronze
12967	2213.100315	3005.644431	3719.771941	Platinum	Platinum	Platinum
12970	4.4711E-06	8.08605E-06	1.35671E-05	Bronze	Bronze	Bronze
12971	2.54001E-69	4.38363E-69	6.87521E-69	Bronze	Bronze	Bronze

Therefore, we have predicted CLTV for 3, 6 and 12 months. Using these CLTV values we segment them using the same method that we used for RFM.

Segmentation using CLTV values (t=3, 6 & 12 months)

Platinum	724
Gold	723
Silver	723
Bronze	723

Each segment contains equal number of customers since we have given equal weightage to each loyalty level.

Loyalty levels according to their average CLTV (in pounds) values is shown below:

Loyalty Levels	clv_3_months	clv_6_months	clv_12_months
Bronze	2.99712E-06	5.2595E-06	8.51572E-06
Silver	0.265449471	0.468103702	0.766016144
Gold	20.93467138	36.90379957	60.54446758
Platinum	342.424011	550.8053904	820.2948913

Conclusion

Constant improvements in data analytics have ensured that the practical applications of models like RFM are seemingly endless. The RFM model ensures effective marketing practices in a world where creating a customer-centric experience is of utmost importance.

The RFM model, when used in conjunction with traditional models of segmentation, can help businesses visualize new and existing customers differently, and create favourable conditions to maximize customer lifetime value. Finding the right balance between focusing on new and existing customers, along with recognizing behavioural nuances within them, will help businesses create personalized customization, leading to brand trust and loyalty.

Customer lifetime value is a relevant metric for any business activity. The difficulty faced by firms in measuring CLTV is the choice of the proper model which offers a satisfactory prediction of customer purchasing behaviour. In a non-contractual relationship, the BG/NBD model is a powerful model to predict customer's purchasing behaviour.

Using R and F scores we have segmented customers into 10 categories, this will help the marketing department to form their specific strategies tailor-made for each customer. Using RFM composite score and future CLTV scores, we can form loyalty levels and marketing team can focus on those customers and give them deals, offers and discounts accordingly.

Limitations

1. The data set we used contained negative values (signifying refunds/returns), the effect of negative values on CLTV is a relatively unknown phenomenon. In order to avoid dealing with that, the positive counterpart of the negative value should be identified and both values should be removed. But, the process of doing that is complicated and time consuming.
2. The segmentation method used in this study was a very basic method, dividing the customers in 5 equal parts, based on their composite score (R+F+M). Machine learning techniques may provide better insights into customer behaviours.
3. Predictions after one period are not accurate, i.e. If we have data for one year, predictions after one year are not very accurate.
4. We assume a Poisson process for the number of transactions, this a poor assumption for seasonal businesses like the one we are studying.
5. In our predictive model, we only include customers with frequency greater than one. This brings down our unique customer count from 4312 to 2893.

References

1. Jo-Ting Wei, Shih-Yen Lin and Hsin-Hung Wu (2010). “A review of the application of RFM model” *African Journal of Business Management* Vol. 4(19), pp. 4199-4206, December Special Review.
2. Fader, P.S., Hardie B.G.S., Lee K.L. (2005) "Counting Your Customers” the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science*, 275-284.
3. Fader, P.S., Hardie B.G.S. (2013) The Gamma-Gamma Model of Monetary Value.
4. Fader P.S., Hardie B.G.S., Lee K.L. (2005). RFM and CLV: Using Iso-Value curves for customer base analysis. J. Mark. Res. 415-430.
5. Lumsden S.A., Beldona S., Morison A.M. (2008). Customer value in an all-inclusive travel vacation club: An application of the RFM framework. J. Hosp. Leisure Mark., 16(3): 270-285.
6. Code and data reference: https://www.kaggle.com/bahaulug/CLTV-prediction-with-bg-nbd-gg-model/data