SGB Data Mining based on Classical Customer Segmentation

Ali Moh

1. Data

The flowing query was used to extract data from database:

```
CREATE OR REPLACE VIEW customers AS
SELECT s.customer_id 'Customer_ID'
, datediff(max(s.updated_at),min(s.created_at)) 'Customer_Life'
, 100/(1+datediff(current_date(),max(s.updated_at))) 'Customer_Recency_Factor'
, sum(s.total_item_count) 'Total_Item_Count'
, count(s.entity_id) 'Total_Orders'
, sum(s.weight) 'Total_Weight'
, sum(s.total_qty_ordered) 'total_qty_ordered'
, sum(s.total_paid) 'total_paid'
, sum(s.total_invoiced) 'total_invoiced'
, sum(datediff(s.updated_at,s.created_at))/count(s.entity_id) 'Average_Order_Duration'
, sum(s.total_refunded) 'Total_Refunded'
, sum(s.total_canceled) 'total_canceled'
, sum(s.shipping_amount)/sum(s.total_item_count) 'shipping_per_item'
from sales_flat_order s
where s.customer_id is not null
group by s.customer_id;
```

It uses sales_flat_order table to extract criteria which their formula can be inferred from the query. The query introduces several new criteria to be used later in customer segmentation algorithm. They are enumerated in the following table which also indicates customer's behavioral factors affected by every criterion;

Criteria			Behavioral Factors				
No.	Name	+/-	Recency	Frequency	Monetary	Breadth	Tenure
1	Customer_Life	+					X
2	Customer_Recency_Factor	+	X				
3	Total_Item_Count	+		X		X	
4	Total_orders	+		X			
5	Total_Weight	+				X	
6	Total_qty_ordered	+		X		X	
7	total_paid	+			X		
8	Total_invoiced	+			X		
9	Average_Order_Duration	-	X			X	
10	Total_Refunded	-			X	X	
11	Total_canceled	-			X	X	
12	Shiping_per_item	-			X		

Column +/- expresses the desired performance of criteria. A positive criterion should be maximized and a negative one should be minimized. Based on above table and sales_flat_order data, the customer segmentation method would be a classical customer segmentation.

2. Algorithm

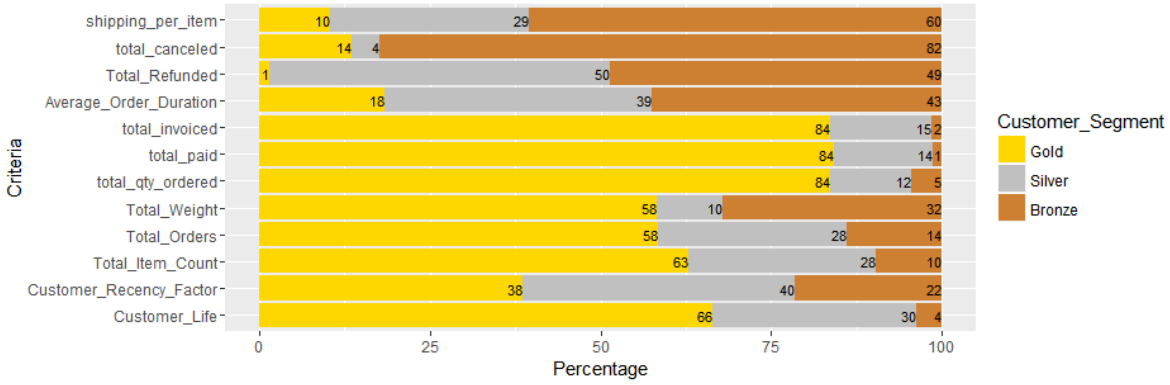
Several Multi-Criteria Decision Making (MCDM) algorithm have been used to rank the customers based on the above mentioned criteria [1, 2]. We choose the most proper one based on customer segmentation criteria. We have a list of ranked customers. Then, based on Pareto principle, we segment the customers in the following three group.

- Gold customers: 20 percent of customers who are top ranked.
- Silver customers: the next 40 percent of customers who are ranked below gold customer and above bronze customers.
- Bronze customers: the next 40 percent of customers who have the lowest ranks.

Gold customers are the best customers who may produce 80 percent of every positive criterion. They should also have the least effect in negative criteria. They should be counted as target customers in any advertisement and promotion investment. A list of segments including customer id and other information is available in Excel format.

3. Analysis

The following figure shows percentage of each criteria occupied by members of customer segments. Successfully, gold customers have maximized values for all positive criteria and they also have minimized values for all negative criteria. It is also good to mention that positive monetary criteria are all above 80 percent for the gold customers. It means gold customers are those 20 percent of customers who provide more than 80 percent of company revenue. Their behavior are also optimized in relation to other criteria.



4. Conclusions

Now, we have a list of ranked customers in three segment; Gold, Silver, Bronze. It is wealthy if we consider gold customers when there is an investment for advertisement and promotions. They are valuable customers who are determined to buy more products. Moreover, they act reasonably fast to complete the transactions which might not be canceled or refunded. Silver customers should be treated as prospect gold customers. They should be tempted in the way that do at least a transaction in a predetermined duration. Bronze customers should be disregarded in any cost effective activity of company.

Codes to produce ranked list of customers and segments can be deployed either as a web service or as a desktop app. It is useful to advise me in this regard.

The next step would be customer clustering which is a modern machine learning based method which may find clusters (segments) of customer based on the factors and features which are hidden in data.

5. References

1. Köksalan, M.M., J. Wallenius, and S. Zionts, *Multiple Criteria Decision Making: From Early History to the 21st Century*. 2011: World Scientific.
2. https://en.wikipedia.org/wiki/Multiple-criteria_decision_analysis. Available from: https://en.wikipedia.org/wiki/Multiple-criteria_decision_analysis.