Customer Segmentation

Kara Akhila
Computer Science
PES University
Banglore, India
karaakhila@gmail.com

Nishtha Varshney
Computer Science
PES University
Banglore, India
nishuvr991@gmail.com

Maknoor Shalini
Computer Science
PES University
Bangalore,India
mshalini0408@gmail.com

Abstract*—Customer Segmentation is a method of partitioning customers into bunches that share comparative attributes. The data-set chosen is extracted from the Machine learning repository. This implementation uses many different approaches to classify customers into other categories-Frequent customers, Infrequent customers, Good customers, low-value customers, high-value customers, champion customers-after preprocessing the extracted data. Analysis of different classifiers used and their pros and cons are presented in this paper. The basic role of this analysis is to enable the business to more readily comprehend its customers and subsequently direct customer driven showcasing all the more viably.

Keywords-segmentation,classification,characteristics,comparing classifiers .

I. INTRODUCTION

In the not so distant past, online retailers centered their showcasing endeavors around engaging 'normal' customers and attracting as many site visitors to their stores as possible. Customers are more modern by they way they explore their shopping decisions, and online retailers are finding that one-size-fits-all promoting techniques aren't helpful any longer indeed focusing on some unacceptable customers can cost you in squandered advertising dollars as well as in higher operational expenses related with preparing item returns, taking care of customer care calls, and reacting to customer surveys. Alternately focusing on the correct customers can take care of you as far as higher normal request esteems and expansion in benefits. Focusing on the correct customers can likewise prompt brand backing and verbal publicizing, important item experiences and more prominent generally speaking consumer loyalty.

One of the greatest tests with customer division is information quality. Off base information in source frameworks will normally give low gathering. Information quality issues additionally emerge from absence of support and customary cleaning to guarantee accuracy.

Different segments of this paper are coordinated as follows.Section 2 discusses related work in this field. Section 3 specifies the problem statement and the data-set description. Section 4 discusses the evaluation metrics and how the classifiers were tested. Section 5 shows experiments and results. Finally, Section 6 concludes the paper.

II. LITERATURE SURVEY

In marketing one approach to build benefits is to speak with customers to decide customer wishes[1]. Correspondence is worked by the attributes of the customer.Customer division requires customer information from different assets. The information is classified into internal and external information[2]. Customer enrollment, client profile and buy history are internal information. Statistics information, media perusing, customer way of life are external information. Customer segmentation can be performed utilizing different methodologies. Hypothetically, Schneider isolates customer division strategies into geographic, segment, psychographic[3]. Baer segments customers utilizing business rule technique, quantile membership method,supervised clustering with classification and unsupervised clustering using the k-means algorithm[1].

Paper	Method	Data	Advantage	Disadvantage
Magento (2014)	Magento	Demographic, Purchase History, Data Product, Data Media, Data Marketing, Server Log	Have clear variable customer segmentation	There is no data processing for each variable
Baer (2012)	Bussiness Rule	Demographic, Purchase history	Easy to apply, Use database query	Not focus on customer behavior
	Quantile membership	Purchase history	Can process small data, can be used with other data	Good result obtained when determining a good classification
	Supervised Clustering with decision tree	Demographic, Purchase history	Classify customers according to target	Use one variable to cluster
	Unsupervised Clustering	Purchase history	Use any number of customer attributes	Speed of computation depends on k values
Colica (2011)	Customer Profiling	Demographic, Purchase history	use database query if data is small	Not focus on behavior
	Customer Likeness Clustering	Demographic, Purchase History, Data product	classify customers according to the target	Problem arises when there are different unit in record
	RFM Cell Classification Grouping	Purchase history	Efficient three -dimensional mapping	Good result obtained when determining a good classification
	Purchase Affinity Clustering	Purchase history, Data product	know the products most in demand	Spesific to product segmentation

[5] Grouping customers merely based on their expenses may not yield desired outcomes. A model which considers more parameters into consideration is more reliable. One of such methods is the RFM analysis.It scores the customers based on three factors.

- R-Recency
- F-Frequency
- M-Monetary

We can then apply any clustering models such as k-means clustering to segment our customers with similar behaviours detected through RFM scoring.

With the headway of customer arranged conduct in business, creating thought has been paid to customers and their necessities as one of the imperative elements to increase higher profit in the business. Customer relationship management(CRM) hopes to separate customer needs and energize coordinated effort among customers and associations. [4]Late investigations recommend that information changes over the long run and in this way the outcomes would not be useful (Roddick and Spiliopoulou, 2002). The current world is in a consistent state of flux that makes the previous results obsolete. The new philosophy is to mine these movements after some time periods. Chen et al (2005) attempted to join customer conduct factors, segment factors and exchange information bases to introduce a technique for mining changes in customer conduct. In 2009 Böttcher introduced an arrangement of customer division dependent on the disclosure of incessant itemsets and the examination of their change after some time. Thus, there is an absence of studies on customer regard assessment dependent on bunching procedures.

III. PROBLEM STATEMENT AND DATA-SET DESCRIPTION

Separating the customers into gatherings of people that are comparable in explicit manners pertinent to advertising, for example, age, gender, interests and ways of managing money. How can we divide customers into groups based on the characteristics they share ? We will discuss various approaches in further sections.
The link for the chosen data-set is given below:
https://archive.ics.uci.edu/ml/datasets/Online+Retail#

This is a value-based informational dataset containing all the transactions for the year 2010-11 for a United Kingdom based online retail store. The dataset consists of 8 attributes.

Attribute information:

- invoice number: ID of the transaction.
- stock code : Product stock ID
- description : Product details
- quantity : Number of products purchased
- invoicedate : Date and time of the transaction
- unit price : Price of the product
- CustomerID : unique identification of the customer
- Country : the country of purchase.

IV. ANALYSIS

Sales and marketing supervisors need to fragment their clients

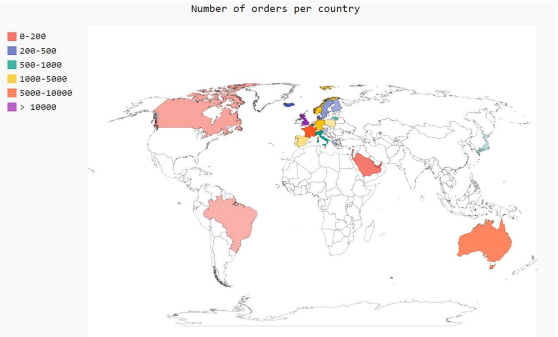
and information so as to offer leads, openings, and past clients customized encounters that lead to higher transformations and a superior in general customer experience. Every business is different and every customer is different, so one segment will never cover all the aspects. So to overcome this issue we will be combining all the segments - Behavioral, Demographic, Psychographic and Hybrid segments including the cohort analysis and then make clusters according to their RFM scores in order to get better and focused results

Preprocessing

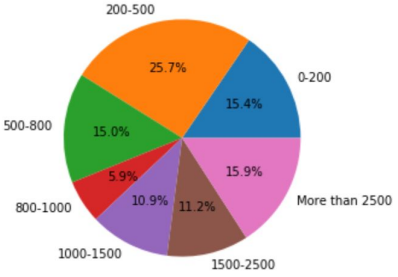
- Missing Values are removed
- Duplicate Values are removed
- There were some negative values observed in the column 'Quantity', which could be due to any of these reasons - it might be purchased and returned product , discount or error. The values which do not have any previous purchase history nor have description as 'Discount' are error values and such rows should be removed from the dataframe.

EDA and Visualisation

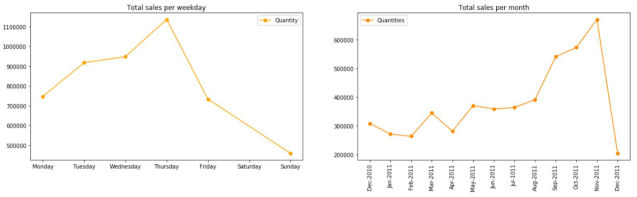
- Dividing the customers based on geographic information.



- Customers and their expenses



- Sales(per weekday/per month)



The above visualisations yield interesting insights:

1. Thursday seems to be the day with the highest number of sales.
2. Friday and Sunday have very less sales.
3. The Pre-Christmas season starts in september due to which there's a peak in sales in november
4. February and April have very low sales

V. Proposed Approach

1. Cohort Analysis(Time based cohorts)

Cohort analysis is a tool to quantify customer commitment after some time. It assists with knowing whether customer commitment is really improving after some time or is just seeming to improve as a result of development.

Cohort Value Table:

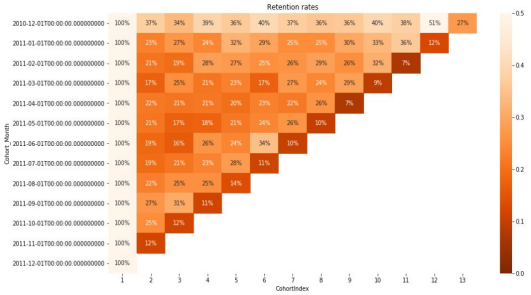
CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
Cohort_Month													
2010-12-01	885.0	331.0	297.0	344.0	322.0	354.0	328.0	315.0	321.0	351.0	332.0	450.0	243.0
2011-01-01	417.0	96.0	112.0	99.0	134.0	123.0	104.0	104.0	126.0	138.0	152.0	52.0	NaN
2011-02-01	380.0	78.0	74.0	108.0	104.0	96.0	99.0	109.0	97.0	120.0	27.0	NaN	NaN
2011-03-01	452.0	78.0	115.0	94.0	103.0	78.0	123.0	108.0	130.0	41.0	NaN	NaN	NaN
2011-04-01	300.0	65.0	62.0	64.0	60.0	69.0	65.0	78.0	22.0	NaN	NaN	NaN	NaN
2011-05-01	284.0	59.0	49.0	50.0	61.0	67.0	75.0	28.0	NaN	NaN	NaN	NaN	NaN
2011-06-01	242.0	45.0	39.0	64.0	57.0	83.0	25.0	NaN	NaN	NaN	NaN	NaN	NaN
2011-07-01	188.0	36.0	39.0	44.0	52.0	21.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08-01	169.0	38.0	43.0	43.0	23.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09-01	299.0	80.0	92.0	34.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10-01	358.0	89.0	42.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11-01	324.0	38.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12-01	41.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Retention Rate Table:

Retention gives the percentage of active customers compared to the total number of customers.

CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
Cohort_Month													
2010-12-01	100.0	37.4	33.6	38.9	36.4	40.0	37.1	35.6	36.3	39.7	37.5	50.8	27.5
2011-01-01	100.0	23.0	26.9	23.7	32.1	29.5	24.9	24.9	30.2	33.1	36.5	12.5	NaN
2011-02-01	100.0	20.5	19.5	28.4	27.4	25.3	26.1	28.7	25.5	31.6	7.1	NaN	NaN
2011-03-01	100.0	17.3	25.4	20.8	22.8	17.3	27.2	23.9	28.8	9.1	NaN	NaN	NaN
2011-04-01	100.0	21.7	20.7	21.3	20.0	23.0	21.7	26.0	7.3	NaN	NaN	NaN	NaN
2011-05-01	100.0	20.8	17.3	17.6	21.5	23.6	26.4	9.9	NaN	NaN	NaN	NaN	NaN
2011-06-01	100.0	18.6	16.1	26.4	23.6	34.3	10.3	NaN	NaN	NaN	NaN	NaN	NaN
2011-07-01	100.0	19.1	20.7	23.4	27.7	11.2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08-01	100.0	22.5	25.4	25.4	13.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09-01	100.0	26.8	30.8	11.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10-01	100.0	24.9	11.7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11-01	100.0	11.7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12-01	100.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Heat Map :



2. Clustering by means of RFM Scores:

Customers are grouped and analysed based on their RFM scores. RFM scores are calculated for each customer based on their past purchases.

Recency score(R) **:**

R score is determined based on the customer's latest purchase. The more recent the last purchase ,the greater the R score.

```
r_labels = range(5, 0, -1)
r_groups = pd.qcut(dfrfm['rvalue'], q=5, labels=r_labels)
dfrfm = dfrfm.assign(R = r_groups.values)
dfrfm
```

	custid	rvalue	R
0	17850.0	3537.997352	1
1	13047.0	3267.100130	4
2	12583.0	3238.268880	5
3	13748.0	3331.200825	2
4	15100.0	3569.166102	1
...	...	...	...
4334	13436.0	3237.167569	5
4335	15520.0	3237.150208	5
4336	13298.0	3237.057847	5
4337	14569.0	3236.983541	5
4338	12713.0	3236.096041	5

4339 rows × 3 columns

Frequency score(F) :

F score indicates the number of purchases the customer has made over a given period of time.

```
f_labels = range(1, 6)
f_groups = pd.qcut(dfrfm['freq'], q=5, labels=f_labels)
dfrfm = dfrfm.assign(F = f_groups.values)
dfrfm
```

	custid	rvalue	R	freq	F
0	17850.0	3537.997352	1	300	5
1	13047.0	3267.100130	4	175	5
2	12583.0	3238.268880	5	248	5
3	13748.0	3331.200825	2	28	2
4	15100.0	3569.166102	1	3	1
...	...	...	...	...	...
4334	13436.0	3237.167569	5	12	1
4335	15520.0	3237.150208	5	18	2
4336	13298.0	3237.057847	5	2	1
4337	14569.0	3236.983541	5	12	1
4338	12713.0	3236.096041	5	38	3

4339 rows × 5 columns

Monetary Score(M) :

M score is determined by the total money spent by the customer. More the money spent, greater is the M score.

```
m_labels = range(1, 6)
m_groups = pd.qcut(dfrfm['mvalue'], q=5, labels=m_labels)
dfrfm = dfrfm.assign(M = m_groups.values)
dfrfm
```

	custid	rvalue	R	freq	F	mvalue	M
0	17850.0	3537.997352	1	300	5	5344.35	5
1	13047.0	3267.100130	4	175	5	3196.09	5
2	12583.0	3238.268880	5	248	5	7220.54	5
3	13748.0	3331.200825	2	28	2	948.25	4
4	15100.0	3569.166102	1	3	1	876.00	3
...	...	...	...	...	...	...	...
4334	13436.0	3237.167569	5	12	1	196.89	1
4335	15520.0	3237.150208	5	18	2	343.50	2
4336	13298.0	3237.057847	5	2	1	360.00	2
4337	14569.0	3236.983541	5	12	1	227.39	1
4338	12713.0	3236.096041	5	38	3	848.55	3

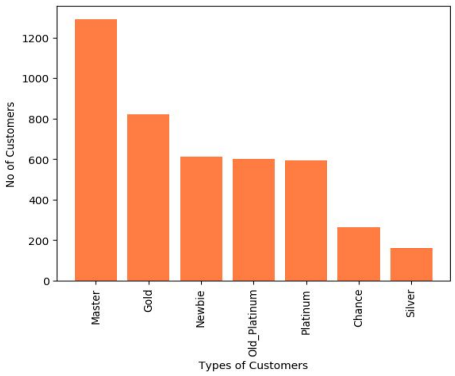
4339 rows × 7 columns

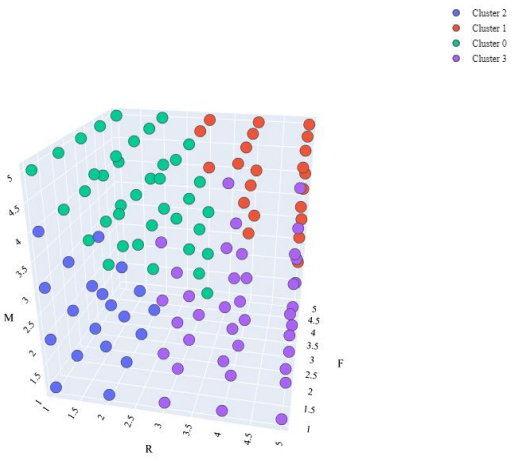
Now that the RFM scores are determined, the next step is to group the customers according to our requirements .

RFM values set according to our preferences:

1. Newbies: customers with high R scores but low F and M scores.
2. Silver: customers with high R and M scores and less F score.
3. Gold: customers with high R and F but low M score.
4. Platinum: customers with high R,F,M scores.
5. Old_platinum : customers with low R scores but high F and M scores.

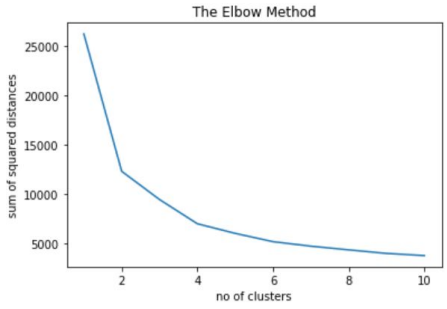
6. Chance: customers with low R and F scores but high M scores.





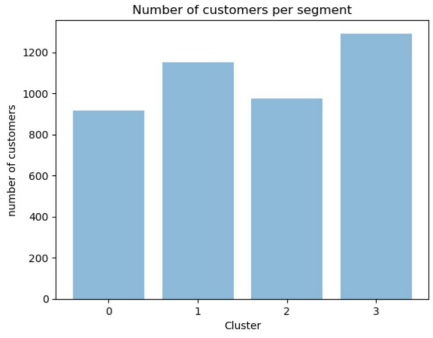
K-means Clustering:

The Elbow method helps us in determining the optimistic number of clusters.



According to the graph k=4.

Number of customers per each cluster are as shown



Plotting the clusters:

Calculating average R,F,M scores for each cluster.

Cluster 1 has an elite of customers with high R,F,M scores, whereas cluster 0 has customers with low R,F,M values. Customers in cluster 2 are frequent and make high valued purchases but they have not made any purchase recently. And customers in cluster 3 have high R score but they are neither frequent nor do they spend more money on their orders.

VI. Conclusion

Customer Segmentation is the first task to make an effective showcasing technique and furthermore the foundation of customer relationship management. The cluster analysis of the chosen dataset explained a lot about the possible clusters in the target population of customers. We started with the analysis of the customer behaviour using cohort analysis, to know the customer commitment and then made clusters with RFM scores. RFM scores are calculated for each customer based on the past purchases. Once the number of clusters were determined using the elbow method, k-means clustering was used to make the clusters of the customers. Cluster 2 - old_platinum customers are frequent and make high valued purchases but they have not made any purchase recently, Cluster 0 - chance customers (low RFM scores), Cluster 3 - Newbies have high R score but they are neither frequent nor do they spend more money on their orders and Cluster 1 - platinum customers (high RFM scores). This

practice of segmenting the customers will help in delivering a better customer experience by focusing on their needs and wants and improving the marketing investments by targeting the only those prone to the best clients

References

[1] Baer D. CSI : Customer Segmentation Intelligence forIncreasing Profits.SASGlobForum.2012:1-13.http://support.sas.com/resources/papers/proceedings12/103-2012.pdf.

[2] Magento. An Introduction to Customer Segmentation. 2014.info2.magento.com/.../An_Introduction_to_Customer_Segmentation...

[3] Schneider G. Electronic Commerce, 9th Edition.; 2013:643. doi:10.1002/1521-3773(20010316)40:6<9823::AIDANIE9823>3.3.C

[4] https://www.researchgate.net/publication/282942257_New_approach_to_customer_segmentation_based_on_changes_in_customer_value

[5] International Journal of Contemporary Economics and Administrative Sciences