Data-driven Segmentation of Consumers' Purchase Behaviour in the Retail Industry

George Carmichael, Yu-wang Chen *
Alliance Manchester Business School
The University of Manchester
Manchester M13 9SS, UK
e-mail: yu-wang.chen@manchester.ac.uk

Cheng Luo
International Business School
Zhejiang Gongshang University
Hangzhou, Zhejiang
e-mail: luochenggemini@hotmail.com

Abstract—**In the modern world, traditional marketing approaches are gradually being dropped in favour of data-driven business analytics due to the improved efficiency and consumer relevancy. Customer segmentation is leading the way in marketing-based analytics as it allows consumers to be grouped by their purchase behaviours. This research presents the application of a customer segmentation model in a dynamic and competitive American retail market, with the aim of producing insightful marketing strategies. 994 consumers with 168,621 transactions were monitored over a four-year period using the IRI store scanner data. Each consumer was segmented using cluster analysis to four segments, namely (1) 'Bargain hunters', (2) 'Opportunistic explorers', (3) 'Promotion averse exploiters', and (4) 'Opportunistic exploiters', were identified in the US salt-snack market.**

Keywords-Customer segmentation; clustering; purchase behaviour

I. Introduction

Since the turn of the 21st century, the retail market has become increasingly competitive and refined. Resultantly companies are searching for innovative developments which will give them an advantage, enabling differentiation from rivals. A key area where this differentiation can occur is marketing. It is no longer effective to target all customers with the same marketing content (Rossi, et al., 1996). The risk of alienating customers with irrelevant content is high and a lack of personalised marketing is unlikely to attract customers, nor increase sales (Smith & Cooper-Martin, 1997). Customer segmentation allows specific groups of people to be identified and targeted which can lead to significant improvements in marketing campaigns and increased business intelligence. This segmentation of customers provides the opportunity to genuinely understand consumers and build models that can predict their behaviour. As noted by Keegan and Green (2008), segmentation of customers therefore allows the number of selling opportunities to be maximised whilst still considering the limited availability of resources. Customer segments are valuable as they provide improved selling opportunities for companies who must efficiently use their limited resources for marketing purposes (Beane & Ennis, 1987). This focus gives organisations a potential advantage in the business world as marketing costs can be reduced at the same time as reaching more consumers with content that is both relevant and persuasive. Customer segmentation is a technique studied deeply in academic literature (Beane and Ennis, 2007; Chan, 2008; Heilman et al., 2000; Lichtenstein et al., 1997). For example, prior studies have segmented customers based upon promotion proneness and brand selection using past data. Whilst some previous studies have been based on real historical purchase data (Heilman et al., 2000; Yoon and Tran, 2011), others have relied upon self-reported data such as interviews (Lichtenstein et al., 1997; Lin, 2002). Although these studies considered the impact of promotions on consumer purchase behaviour, they did not analyse the trade-off between market exploration and promotion seeking behaviours. Furthermore, the earlier studies relied on more traditional methods of segmentation such as visualisation and self-reported data, which are now considered less reliable than the modern methods which utilise powerful predictive models (Witten, et al., 2011).

This research builds on prior research by considering the trade-off between prevalence of promotion and value of information from purchases using advanced modelling capabilities. This is achieved through customer segmentation based on two computed inputs that measure these attributes. Marketing recommendations are then provided based upon the findings of this research. The rest of the paper is organised as follows. In Section II, the measures of purchase behaviour used for segmentation are outlined and evaluated. Section III discusses the analysis steps including the reasoning behind the methodology. Section IV outlines the results of the project with references to its real-world implications and finally, and Section V provides a conclusion which includes the marketing recommendations and the limitations of the project.

II. Behavioural Measurements

In this section, two variables with regards to maximising immediate purchase value and market exploration behaviours respectively, namely 'prevalence of promotion' and 'value of information' are introduced as inputs for behavioural segmentation.

A. *Prevalence of Promotion*

Promotion proneness is usually defined as the extent to which a consumer is motivated to search and take advantage of promotions to maximise the immediate purchase value. The attitudes and tastes towards promotions vary amongst consumers and thus their reaction to a promotion will also

978-1-5386-6147-5/18/$31.00 ©2018 IEEE

differ (Lichtenstein, et al., 1993). A customer who frequently buys products on promotion is said to be 'promotion prone' whereas a customer who is not tempted by promotion and resultantly tends to buy products at full price is described as 'promotion averse' (DelVecchio, 2005). This measure can simply be calculated as follows

$$\textit{Prevalence of promotion} = p \mathbin{/} P \tag{1}$$

where p denotes the total number of purchases on promotion in a period, while P is the total number of ~~promotions~~ in the period. Therefore, the more purchases made on promotion relative to the total number of purchases, the higher the prevalence of promotion metric is (Anic and Radas, 2006). Accordingly, marketers can use the measure to target customers who are more likely to respond positively to promotions to save costs and improve the efficiency of an advertising campaign.

B. *Value of Information*

The brand selection behaviours of consumers change as new information is collected through the purchasing of unfamiliar brands. This reduces the perceived risks of trying new products (Heilman et al., 2000). This reduced uncertainty about the market from trying new goods is described as the "Value of Information from Purchases" (Luo et al., 2015). The value of information is a function heavily studied in information theory for application in financial markets (Chen, 2004). It is a function of probability known as the generalised entropy measurement. In this paper, this metric is used to quantify the dynamic choice behaviours of consumers in the US salt-snack market. The value of information input is based upon the market knowledge of the consumer. Formula 2 shows the calculation of market knowledge which forms the basis of the value of information from purchase computation.

$$\textit{Market Knowledge} = n \mathbin{/} N \tag{2}$$

where n represents the number of brands tried by a consumer in their purchase life cycle, while N is the total number of brands available in the market during the consumer's purchase life cycle. Both the obtainable value of information and market knowledge levels of the consumer must be considered in the final calculation (Heilman et al., 2000). Therefore, it can be stated that the true value of information from purchases (the reduction of risks due to increased market knowledge) can be quantified below, which has been adapted from the generalised entropy measurement (Chen, 2004).

$$\textit{Value of information} = (n \mathbin{/} N) \text{ x } (-\log_2 (n \mathbin{/} N)) \tag{3}$$

This calculation forms an Iiverted-U shaped relationship between market knowledge and the value of information measures. As market knowledge increases from zero, so does the value of information from purchases. However, once market knowledge is sufficient (around 0.5), the value of information begins to drop. This process reflects the consumer's mind-set with regards to risk and uncertainty (Heilman et al., 2000); a customer who is new to a market has little motivation to explore a market as they are unable to differentiate between product attributes, so the purchase risk is high. Once the consumer has tried a few of the major brands their market knowledge increases enough for them to feel more confident in trying new brands. They then explore the market further in search of an 'ideal' brand. Midway through this process their motivation to explore the market begins to drop as the customer realizes that they have tried most brands and hence, have likely already found the brand that most meets their needs and wants. Value of information hence reflects the motivation of a consumer to gain new information through brand exploration for any given level of market knowledge. A high value of information from purchases (close to 1) indicates a strong motivation to try new brands to extend market knowledge. Consumers in this position are categorised as 'explorers'. In contrast, consumers with a low value of information from purchases (close to 0) are more likely to stick to familiar brands to avoid taking risks. These consumers are classified as 'exploiters' as they have little motivation to try new brands.

III. Data-driven Segmentation of Consumers' Purchase Behaviour

A. *IRI Dataset*

The IRI dataset is widely used in marketing research and is a very valuable source of information (Bronnenberg et al., 2008). Not only is it very accessible to academics but the demographic information for customers is very comprehensive (26 demographic variables each year), thus creating significant research opportunities. In addition to this, it has more complete purchase information than typical loyalty card data as it is self-recorded data, carried out by the consumer after a shopping trip (Bronnenberg, et al., 2008). For this reason, it may also be more accurate than loyalty card data as it does not rely on consumers remembering to bring or use a loyalty card to the store. These benefits suggest the IRI dataset can be considered a suitable data source for academic data mining and analysis purposes.

This research was based upon customers who shop in the Eau Claire market as 6 retailers provided transactional data to IRI which is a high participation level, relative to other locations in the IRI datasets. Eau Claire is small city located in the state of Wisconsin in the United States. In 2010 it had a population of 65,883 and has seen population growth of 7% or more each year over the last 3 decades. This market thus offers high potential for maximising future selling opportunities due to its growing customer base. The model studied in this paper is only applied to transactions from 2004 to 2007 to ensure demographic comparisons across years. The salt-snack market was chosen as it is a very turbulent market where products are bought frequently meaning there is not only a large supply of continuous data over the years, but also that it has many real-world applications in retail markets.

B. *Data Processing*

In the Eau Claire market the salt-snack industry is very competitive. In 2004 there were 79 brands available to consumers and in the following three years 8, 11 and 8 new brands entered the market respectively. This resulted in consumers having a total of 106 brands to choose from over the four-year period used for analysis. Over the four years the average market knowledge level was only 9% and no consumers tried more than 23% of the brands during the four-year period. These figures show the competitiveness of the salt-snack market in Eau Claire, with over 100 brands on offer it is very difficult for consumers to fully explore the market due it's dynamic growth and competitive nature.

- Customer Filtering

Between 2004 and 2007 in the Eau Claire grocery market, there were 4029 customers who made at last one purchase during the four years and 1726 (43%) of those made a purchase at least once every year. To quantify prevalence of promotion and value of information behaviours, a sufficient number of transactions are needed per customer for each of the four years. This is important to ensure the behavioural measurements used for segmentation are reliable and based on an adequate volume of purchases each year. If the number of sales observations per year per customer is too low, then the analysis results are likely to be unrealistic for a typical customer in the salt-snack industry.

This was an important step of the data processing as a cut-off point for the number of transactions per year had to be set to define an 'appropriate customer' for analysis. A trade-off occurred at this stage as the higher the cut-off point, the more suited the data was for analysis. However, at the same time it reduced the size of the learning and validation datasets for the segmentation model. If a model is based on a very low number of customers it is likely to show bias based on their behaviours and hence be less applicable. After considering these limitations, a cut-off point of 12 purchases per year between 2004 and 2007 was chosen. This created a group of 997 customers who met the transactional requirements in the Eau Claire grocery market.

Two more filters were applied to the data before the variable calculation steps. Firstly, it was important for the behavioral validation of the model, that all customers had complete sets of demographic data to enable demographic profiling of segments. All 997 customers filtered above already had full demographic profiles so no customers were removed from analysis at this stage. Secondly, as the prevalence of promotion measurement relies upon promotional information it was decided that customers could be missing no more than 10% of promotion information from their purchases over the four years. This filter resulted in 3 more customers being removed from the analysis stage, resulting in 994 eligible customers for analysis, sharing 168,621 transactions between them over the four years.

- Learning/Validation Datasets

The customer segmentation model requires a learning and validation dataset. The learning dataset is used as a basis for the model and the validation dataset is to ensure the model is effective on unseen data. In this research a 60:40 split was used to select the learning and validation datasets. Using 60% of data for learning and 40% for validation is a common splitting method backed up by prior research. This led to the learning dataset containing 598 customers and the validation dataset containing 396 customers.

C. *Analysis Process*

The behavioural segmentation was carried out with the purpose of behavioural and demographic profiling as well as analysis of behavioural evolvement patterns over time. Cluster analysis is used in this research as it can deal with complex datasets and finding hidden patterns with the aim of grouping entries in to segments with similar characteristics. In this study, cluster analysis is employed to identify groups of customer with homogenous purchase behaviour. The cluster analysis technique allows the trade-off made by customers, between prevalence of promotion and value of information, to be clearly outlined and identified. Customers with similar behavioural measurements are grouped into the same segment. This implies that customers in the same segment all share similar levels of promotion proneness and brand selection behaviour.

The software used for cluster analysis in this paper is SAS Enterprise Miner as it is a very powerful analysis tool well suited to large and complex datasets. It is widely used by large organisations for its ability to create descriptive and predictive models with high levels of accuracy and versatility. The cluster analysis model was based on the two inputs, prevalence of promotion and value of information, which have been discussed thoroughly in this paper. The two variables had no inter-correlation and it is thus reasonable to assume that four behavioural segments will be identified. *K*-means clustering was employed in the model to set a maximum of four clusters and the initial cluster points were decided by the model.

There was no need to transform the input variables of the cluster analysis as their distributions were already very close to normal and this was confirmed by the max-normalisation node in SAS. Internal standardisation was however used in the cluster analysis node as the values of the two variables differed significantly, with the average prevalence of promotion value being 0.63, whilst the average value of information value was 0.09. The internal standardisation was used to ensure the variables were in a proportionate scale to each other. This ensured the cluster analysis results were reliable and understandable. The scatter graph for prevalence of promotion and value of information prior to analysis showed that the learning and validation datasets had values that were all very close together. For this reason, the full replacement method was used to ensure well separated initial cluster seeds as it repeats the seed replacement test until it has found the optimal cluster centroids based on the seed distances. Although this method is slower than the part-replacement method it ensures a higher level of accuracy which is particularly important here given the lack of separation between clusters. The input variables were set as prevalence of promotion and value on information from the 2004 learning dataset. This behavioural segmentation model was then used to score the remaining seven datasets for the

US salt-snack market. Once the segmentation model had been run, the results of the four segments were analysed to identify the associated purchasing behaviour for each group of customers. Two visual aids, a scatter graph and pair of bar charts, were employed to identify the purchase behaviours of segments.

IV. Results and Discussion

The cluster analysis model produced four segments based on the two inputs of promotion proneness and value of information from purchases. The differences between these behaviour groups are shown in Fig. 1.

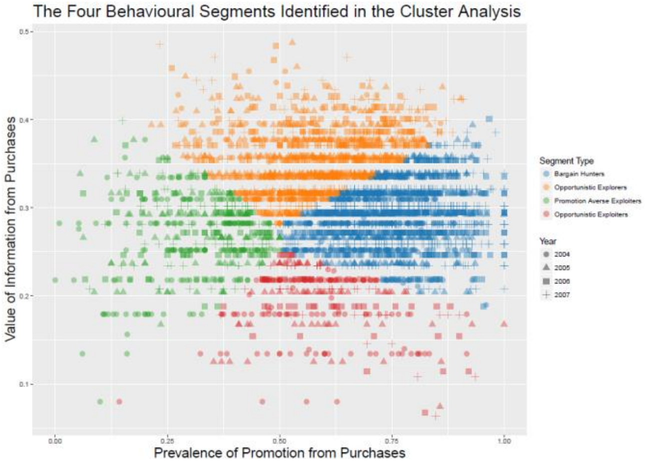

Figure 1. The four segments identified from the cluster analysis model.

The group labelled ‘Bargain hunters’ was found to have high prevalence of promotion levels and medium value of information. The high values for prevalence of promotion show that relative to the total number of purchases by customers in the population, the number of purchases bought on promotion is high for this segment. The average market knowledge in this group was 7% which resulted in the medium levels of value of information from purchases. Consumers in this segment are thus likely to be early on in their purchase life cycle and the medium value of information levels allows this segment to explore the market in search of brands on promotion, with the aim of immediate purchase value maximisation.

The segment labelled “Opportunistic explorers” has high value of information with medium levels of promotion proneness. This high value of information compared the population reflects the segment’s high average market knowledge of value of 11%. These customers have consequently become very good at differentiating between brand attributes and therefore feel they have more control over the risks involved in purchase decisions. Opportunistic explorers thus have a stronger desire to explore the market to find a perfect brand that meets their needs and wants. This segment has the highest value of information and market knowledge in the population with a medium propensity to take advantage of promotions. Therefore, customers in this segment take advantage of promotions to extend their market knowledge.

TABLE I. Profiles of Segments

Segment Name	Prevalence of Promotion	Value of Information	Definition
Bargain Hunters	High (0.75)	Medium (0.29)	Actively looking for promotions to maximise the immediate purchase value.
Opportunistic Explorers	Medium (0.58)	High (0.35)	Motivated to try new brands. Promotions are used to enable this market exploration.
Promotion Averse Exploiters	Low (0.32)	Medium (0.26)	Purchase brands that are well known to them. Unwilling to risk new brands regardless of promotions.
Opportunistic Exploiters	Medium (0.6)	Low (0.19)	Purchase items that they are familiar with to reduce the risks involved whilst taking advantages of promotions where possible.

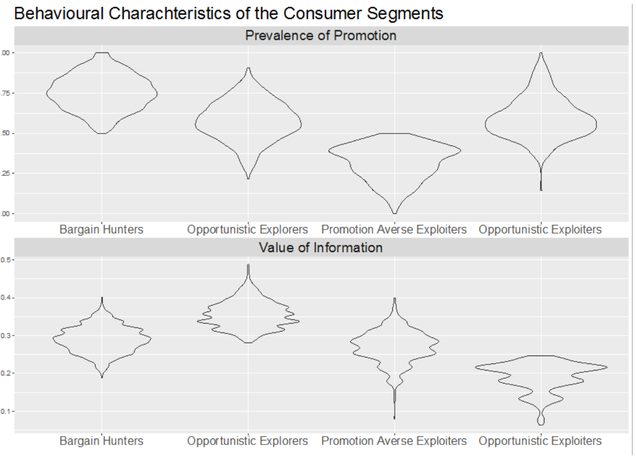

Figure 2. The behavioural charachteristics of each segment.

Consumers in the “Promotion averse exploiters” segment have low levels in prevalence of promotion (average of 0.32 over the four years) therefore showing that consumers in this segments rarely make purchases of salt-snacks that are on promotion. This shows these consumers are unlikely to respond positively to promotions which maybe a result of negative perceptions of brands which are on promotion. This segment has a medium level for value of information with an average of 0.26 over the four years. The segment consumers thus tend to purchase familiar brands regardless of the other promotions with the aim of reducing the risk involved with salt-snack purchases.

“Opportunistic exploiters” have low value of information levels and medium prevalence of promotion from purchases. Customers thus perceive little value in gaining new information about a market through brand exploration. Thus, these customers are very brand loyal with very little incentive to explore the market further. They have little market knowledge (4%) and are hence early in the consumer purchase life cycle. This low value of information is likely a result of customers in the segment struggling to differentiate between products and thus judging the purchase risks as too high to try new brands that they may not like. Instead they choose the safe option of sticking with brands they know and like whilst taking advantage of promotions where possible to minimise risks in salt-snack purchases. The differences between each segment are highlighted in Table 1 and visualised in Fig. 2.

V. Conclusion

This research has been successful in segmenting customers into clear and well-defined behaviour groups. The results in this paper provide valuable findings that can be employed to optimise marketing efforts in an organisation. The customer segmentation model and the associated analysis results have provided insights into customer purchase behaviour, segment demographics and behavioural evolvement. The insights into the four behavioural groups identified in this report can be used to create tailored marketing campaigns based on the target customers. With this information, organisations have a good idea of how effective a campaign will be for each segment and changes can be made to increase the relevance of promotions to each consumer group. Some potential ways of targeting each segment are listed below.

Bargain hunters: This is the easiest group to target as the consumers respond very well to promotions. If marketing budgets are low and high efficiency is important, then this segment should be targeted. Bargain hunters are willing to try new brands if they are on promotion and their main priority is to maximise the immediate purchase value. For this reason, it will be very simple to increase revenue by aiming promotions at this group and it is very likely to cause a large spike in sales due to both their promotion proneness and large segment size.

Opportunistic explorers: Although this segment only has medium promotion proneness, it also has a lot of potential for great marketing campaign results. Opportunistic explorers are very motivated to try new brands and they aim to use promotions to accelerate this process; this can be used to the organisations advantage. If new salt-snack brands are marketed directly at this segment with a small discount, it is very likely to lead to a significant rise in sales. This may help in boosting excitement around a new product launch and establishing a new brand in the salt-snack market.

Promotion averse exploiters: This is the hardest segment to target successfully due to the very low propensity of the customers to purchase salt-snacks that are on promotion. These customers are mainly focussed on purchasing brands they are familiar with, regardless of the promotion status. It may therefore be better to increase sales by offering promotions based around bulk buy offers on the bigger brands they favour. Price reductions on less known products are unlikely to have any impact on this segment and thus should not be targeted with this purpose in mind.

Opportunistic exploiters: This segment is focussed on using promotions to purchase goods they are familiar with. They are very risk averse and it is therefore recommended that promotions on the most popular brands are offered to opportunistic exploiters. This method of marketing will appeal to them and likely lead to good results. The improved immediate purchase value combined with low risk will be very appealing and will likely increase brand loyalty for a store targeting this segment.

The main limitation in this paper has been the data availability with regards to customer transactions. There was a trade-off between having a high number of customers in the segmentation analysis and only considering customers for analysis who have a consistent transactional record over the four years. Customer segmentation provides many opportunities for further analysis which could yield even more valuable insights.

References

[1] Beane, T. & Ennis, D., 1987. Market Segmentation: A Review. *European Journal of Marketing*, 21(5), pp. 20-42.

[2] Bronnenberg, B., Kruger, M. & Mela, C., 2008. The IRI Marketing Data Set. *Marketing Science,* 27(4), pp. 745-748.

[3] Chan, C., 2008. Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer. *Expert Systems with Applications*, 34(4), pp. 2754-2762.

[4] Chen, J., 2004. Generalized entropy theory of information and market patterns. *Corporate Finance Review,* pp. 23-32.

[5] CMO Council, 2013. Online Advertising Performance Outlook, San Jose: CMO.

[6] DelVecchio, D., 2005. Deal-prone consumers' response to promotion: The effects of relative and absolute promotion value. *Psychology of Marketing*, 22(5), pp. 373-391.

[7] Heilman, C., Bowman, D. & Wright, G., 2000. The Evolution of Brand Preferences and Choice Behaviors of Consumers New to a Market. *Journal of Marketing Research,* 37(2), pp. 139-155.

[8] IBM, 2011. *IBM Big Data Platform,* Johannesburg: IBM.

[9] Keegan, W. J. & Green, M. C., 2008. *Global Marketing (5th edition).* Pearson Prentice.

[10] Lichtenstein, D., Ridgway, N. & Netemeyer, R., 1993. Price perceptions and consumer shopping behavior: A field study. *Journal of Marketing Research,* 30(2), pp. 234-245.

[11] Lin, C.-F., 2002. Segmenting customer brand preference: demographic or psychographic. *Journal of Product & Brand Management,* 11(4), pp. 249-268.

[12] Luo, C., de Brujin, O. & Chen, Y.-W., 2015. Behavioural segmentation using store scanner data in retailing: Exploration and exploitation in frequently purchased consumer goods markets. *The Business & Management Review,* 5(3), pp. 43-54.

[13] Rossi, P. E., McCulloch, R. E. & Allenby, G. M., 1996. The Value of Purchase History Data in Target Marketing. *Marketing Science,* 15(4), pp. 321-340.

[14] Smith, C. & Cooper-Martin, E., 1997. Ethics and Target Marketing: The Role of Product Harm and Consumer Vulnerability. *Journal of Marketing,* 61(3), pp. 1-20.

[15] Witten, I., Frank, E. & Hall, M. A., 2011. *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco (California): Morgan Kaufmann.

[16] Yoon, K. & Tran, T., 2011. Capturing consumer heterogeneity in loyalty evolution patterns. *Management Research Review,* 34(6), pp. 649-662.