Technical Article

Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining

Received (in revised form): 18th July 2012

Daqing Chen
is a senior lecturer in the Department of Informatics, Faculty of Business, London South Bank University, London, UK. He mainly lectures in data mining and business intelligence on BSc and MSc courses. His research interests include data mining, data-driven marketing and customer-centric business intelligence. In recent years he has been engaged in several business-oriented data mining projects across various business sectors.

Sai Laing Sain
is currently a BSc student in the Department of Informatics, Faculty of Business, London South Bank University, London, UK.

Kun Guo
is currently a PhD student in the Department of Civil Engineering, Faculty of Engineering, Science, and the Built Environment at London South Bank University, London, UK. His academic interests include numerical modelling, artificial intelligence algorithms and data mining.

ABSTRACT Many small online retailers and new entrants to the online retail sector are keen to practice data mining and consumer-centric marketing in their businesses yet technically lack the necessary knowledge and expertise to do so. In this article a case study of using data mining techniques in customer-centric business intelligence for an online retailer is presented. The main purpose of this analysis is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively. On the basis of the Recency, Frequency, and Monetary model, customers of the business have been segmented into various meaningful groups using the *k*-means clustering algorithm and decision tree induction, and the main characteristics of the consumers in each segment have been clearly identified. Accordingly a set of recommendations is further provided to the business on consumer-centric marketing. SAS Enterprise Guide and SAS Enterprise Miner are used in the present study.

Journal of Database Marketing & Customer Strategy Management (2012) **19,** 197–208.
doi:10.1057/dbm.2012.17; published online 27 August 2012

Keywords: online retail; customer-centric marketing; data mining; customer segmentation; RFM model; *k*-means clustering

Correspondence:
Daqing Chen
Department of Informatics, Faculty of Business, London South Bank University, London, UK
E-mail: chend@lsbu.ac.uk

INTRODUCTION

For the past 10 years, we have witnessed a steady and strong increase of online retail sales. According to the Interactive Media in Retail Group (IMRG), online shoppers in the United Kingdom spent an estimated £50 billion in year 2011, a more than 5000 per cent increase compared with year

2000.[1] This remarkable increase of online sales indicates that the way consumers shop for and use financial services has fundamentally changed.

Compared with traditional shopping in retail stores, online shopping has some unique characteristics: each customer's shopping process and activities can be tracked instantaneously and accurately, each customer's order is usually associated with a delivery address and a billing address, and each customer has an online store account with essential contact and payment information. These desirable, special online shopping characteristics have enabled online retailers to treat each customer as an individual with personalized understanding of each customer and to build upon customer-centric business intelligence.

In relation to customer-centric business intelligence, online retailers are usually concerned with the following common business concerns:

- Which items/products' web pages has a customer visited? How long has a customer stayed with each web page, and in which sequence has a customer visited a set of products' web pages?
- Who are the most/least valuable customers to the business? What are the distinct characteristics of them?
- Who are the most/least loyal customers, and how are they characterized?
- What are customers' purchase behaviour patterns? Which products/items have customers purchased together often? In which sequence the products have been purchased?
- Which types of customers are more likely to respond to a certain promotion mailing? and
- What are the sales patterns in terms of various perspectives such as products/items, regions and time (weekly, monthly, quarterly, yearly and seasonally), and so on?

In order to address these business concerns, data mining techniques have been widely adopted across the online retail sector, coupled with a set of well-known business metrics about customers' profitability and values, for instance, the recency, frequency and monetary (RFM) model,[2] and the customer life value model.[3] For many online retailers in the United Kingdom and internationally alike, especially the leading companies including Amazon, Walmart, Tesco, Sainsbury's, Argos, Marks and Spencer, John Lewis, and EasyJet, data mining has now become a common practice and an integral part of the business processes in creating customer-centric business intelligence and supporting customer-centric marketing.[4,5]

Although many famous online retail brands are embracing data mining techniques as crucial tools to gain competitive advantages on the market, there are still many smaller ones and new entrants are keen to practise consumer-centric marketing yet technically lack the necessary knowledge and expertise to do so.

In this article a case study of using data mining techniques in customer-centric business intelligence for an online retailer is presented. The online retailer considered here is a typical one: a small business and a relatively new entrant to the online retail sector, knowing the growing importance of being analytical in today's online businesses and data mining techniques, however, lacking technical awareness and recourses. The main purpose of this analysis is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively. On the basis of the RFM model, customers of the business have been segmented into various meaningful groups using the *k*-means clustering algorithm and decision tree induction, and the main characteristics of the consumers in each segment have been clearly identified. Accordingly, a set of recommendations is provided to the

business on customer-centric marketing and further data analysis tasks. The analysis is developed in a step-by-step way. SAS Enterprise Guide and SAS Enterprise Miner[6–9] have been employed in this study.

The rest of this article is organized as follows. The next section provides the background information about the online retailer studied in the article along with the associated dataset to be explored. The section after that discusses in detail about the main steps and tasks for data pre-processing in order to create an appropriate target dataset for the required further analyses. In the subsequent section the *k*-means clustering analysis is performed and a set of meaningful clusters and segments of the target dataset has been identified. A detailed discussion on each of the clusters is given, and the segmentation is further refined by using decision tree induction. The penultimate section summarizes the essential consumer-centric business intelligence based on the analysis results, and provides some concrete recommendations to the online retailer aiming at maximizing profits for the business. Finally the concluding remarks are given in the last section.

BUSINESS BACKGROUND AND THE ASSOCIATED DATA

The online retailer under consideration in this article is a UK-based and registered non-store business with some 80 members of staff. The company was established in 1981 mainly selling unique all-occasion gifts. For years in the past, the merchant relied heavily on direct mailing catalogues, and orders were taken over phone calls. It was only 2 years ago that the company launched its own web site and shifted completely to the Web. Since then the company has maintained a steady and healthy number of customers from all parts of the United Kingdom and Europe, and has accumulated a huge amount of data about many customers. The company also uses Amazon.co.uk to market and sell its products.

The customer transaction dataset held by the merchant has 11 variables as shown in Table 1, and it contains all the transactions occurring in years 2010 and 2011. It should be noted that the variable *PostCode* is essential for the business as it provides vital information that makes each individual consumer recognizable and trackable, and therefore it makes some in-depth analyses possible in the present study.

As the first ever pilot study for the business to generate sensible customer intelligence, only the transactions created from 1 January 2011 to 31 December 2011 are explored in this article. Over that particular period, there were 22 190 valid transactions in total, associated with 4381 valid distinct postcodes. Corresponding to

Table 1: Variables in the customer transaction dataset (4381 instances)

Variable name	*Data type*	*Description; typical values and meanings*
Invoice	Nominal	Invoice number; a 6-digit integral number uniquely assigned to each transaction
StockCode	Nominal	Product (item) code; a 5-digit integral number uniquely assigned to each distinct product
Description	Nominal	Product (item) name; CARD I LOVE LONDON
Quantity	Numeric	The quantities of each product (item) per transaction
Price	Numeric	Product price per unit in sterling; £45.23
InvoiceDate	Numeric	The day and time when each transaction was generated; 31/05/2011 15:59
Address Line 1	Nominal	Delivery address line 1; 103 Borough Road
Address Line 2	Nominal	Delivery address line 2; Elephant and Castle
Address Line 3	Nominal	Delivery address line 3; London
PostCode	Nominal	Delivery address postcode, mainly for consumers from the UK; SE1 0AA
Country	Nominal	Delivery address country; England

these transactions, there are 406 830 instances (record rows) in the dataset, each for a particular item contained in a transaction. On average, each postcode is associated with five transactions, that is, each customer has purchased a product from the online retailer about once every 2 months. In addition, only consumers from the United Kingdom are analysed.

It is interesting to notice that the average number of distinct products (items) contained in each transaction occurring in 2011 was 18.3 (= 406 830/22 190). This seems to suggest that many of the consumers of the business were organizational customers rather than individual customers.

DATA PRE-PROCESSING

In order to conduct the required RFM model-based clustering analysis, the original dataset needs to be pre-processed. The main steps and relevant tasks involved in the data preparation are as follows:

1. Select appropriate variables of interest from the given dataset. In our case the following six variables have been chosen: *Invoice, StockCode*, *Quantity*, *Price*, *InvoiceDate* and *PostCode.*
2. Create an aggregated variable named *Amount*, by multiplying *Quantity* with *Price*, which gives the total amount of money spent per product/item in each transaction.
3. Separate the variable *InvoiceDate* into two variables *Date* and *Time*. This allows different transactions created by the same consumer on the same day but at different times to be treated separately.
4. Filter out any transactions that do not have a postcode associated with. This resolves any missing value issues in relation to the variable *PostCode*. In addition, filter out any transactions that are not associated with a United Kingdom's postcode.
5. Sort out the dataset by *Postcode* and create three essential aggregated variables *Recency*, *Frequency* and *Monetary*. Calculate the values of these variables per postcode.

Following these steps a target dataset for the analysis has been generated. The original dataset was in MS Excel format, and was transformed into the final target dataset in SAS format in SAS Enterprise Guide 4.2. Part of the target dataset is shown in Figure 1, and the variables in the target dataset and their statistics are described in Tables 2 and 3. The SAS procedures *proc means* and *proc sql* were used to transform the dataset and to calculate the values for the variables *Recency*, *Frequency* and *Monetary*, for each given postcode, respectively. As an example, Table 4 gives the relevant SAS code utilized to calculate the values for *Monetary*. Finally the target dataset was uploaded into SAS Enterprise Miner 6.2 for analysis.

RFM MODEL-BASED CLUSTERING ANALYSIS

Clustering

With the prepared target dataset we intended to identify whether consumers can be segmented meaningfully in the view of recency, frequency and monetary values. The *k*-means clustering algorithm was employed for this purpose, and it can be easily performed by using the Cluster node in SAS Enterprise Miner.

As well-known, the *k*-means clustering algorithm is very sensitive to a dataset that contains outliers (anomalies) or variables that are of incomparable scales or magnitudes. Examining the histograms of the variables *Recency*, *Frequency* and *Monetary* of the target dataset in SAS Enterprise Miner, as illustrated in Figure 2, it is evident that there are a few instances having quite different monetary and

Filter and Sort | Query Builder | Data | Describe | Graph | Analyze | Export | Send To

	Buyer	First_Purchase	Recency	Frequency	Monetary	Min	Max	Mean
2392	XX18 4ES	12	12	1	306.84	1.25	30	7.14
2393	XX2 4HG	4	4	1	1487.6	237.6	1250	743.80
2394	XX2 5XX	5	0	5	1952.45	0.39	76.32	22.19
2395	XX20 6AB	11	1	2	483.26	0.84	17.34	4.56
2396	XX20 6AQ	7	7	1	157.9	10.5	75	26.32
2397	XX20 6HJ	3	1	3	541.46	1.65	59.4	15.93
2398	XX21 8YN	1	1	1	167.62	0.19	15.12	3.64
2399	XX22 3RB	3	1	3	454.01	2.9	25.5	9.27
2400	XX24 8SA	11	11	1	344.14	1.45	40.56	11.87
2401	XX25 5HU	10	2	4	1498.51	0.12	45	5.01
2402	XX26 1AG	12	3	4	226.75	13.05	26.1	18.90
2403	XX26 3QN	8	8	1	185.65	15	35.7	23.21
2404	XX30 6AP	12	8	3	215.72	0.58	47.7	4.79
2405	XX30 6BJ	8	8	1	175.27	0.65	60	6.04
2406	XX31 7AD	11	4	4	1133.07	0.29	41.5	5.29
2407	XX31 7JF	8	4	3	1240.2	9.12	264	53.92
2408	XX31 7JN	8	2	4	2404.17	0.39	74.25	9.50
2409	XX33 0EA	9	3	2	993.18	8.5	39.6	19.86
2410	XX33 0EN	1	1	2	388.79	0.39	34.68	4.68
2411	XX33 0QL	12	1	7	2827.93	0.29	142.8	10.47
2412	XX34 1HA	7	7	1	2044.37	5.04	207.5	37.17
2413	XX34 2DS	1	1	1	161.67	2.5	19.8	12.44
2414	XX34 2HA	8	8	1	116.01	2.1	30	12.89
2415	XX34 3OA	11	7	2	309.36	1.1	17.4	4.91
2416	XX37 6QU	1	1	2	193.42	0.19	11.8	3.28
2417	XX39 4HL	2	2	1	390.07	0.42	18.72	3.90

Figure 1: Samples of the target dataset.

Table 2: Variables in the target dataset

Variable name	*Data type*	*Description*
Buyer	Nominal	Corresponding to each distinct postcode
Recency	Numeric	Recency in month
First_Purchase	Numeric	Time in month since the first purchase in 2011
Frequency	Numeric	Frequency of purchase per postcode
Monetary	Numeric	Total amount spent per postcode
Minimum	Numeric	Minimum spending per postcode
Maximum	Numeric	Maximum spending per postcode
Mean	Numeric	Median spending per postcode

Table 3: Summary of the target dataset (3799 instances)

Variable name	*Minimum*	*Median*	*Maximum*
Recency	0	3.2	12
Frequency	1	4.9	169
Monetary	3.75	1586.63	88 125.38
First_Purchase	0	7.5	12

Table 4: Sample SAS codes for calculating values of monetary

```
proc means data=YourLibraryName.
   SortedOriginalDataset n sum min
max mean;
var Amount;
by Postcode;
output out=YourLibraryName.TagretDatasetMonetary
(drop=_type_ _freq_) n=n sum=sum min=min max=max
   mean=mean;
run;
```

frequency values compared to the majority of the instances in the dataset. These instances are valid from the business point of view as they are genuine transaction records; however, they are outliers from the data analysis point of view. Therefore, these instances should be isolated from the majority and treated separately. In addition, the three variables are not on comparable scales, and the value ranges are quite different: *Recency* [0,12]; *Frequency* [1,169] and *Monetary* [3,88 125], respectively. As such, these variables should be normalized before the clustering analysis.

On the basis of the initial insight into the dataset, a project diagram has been set up

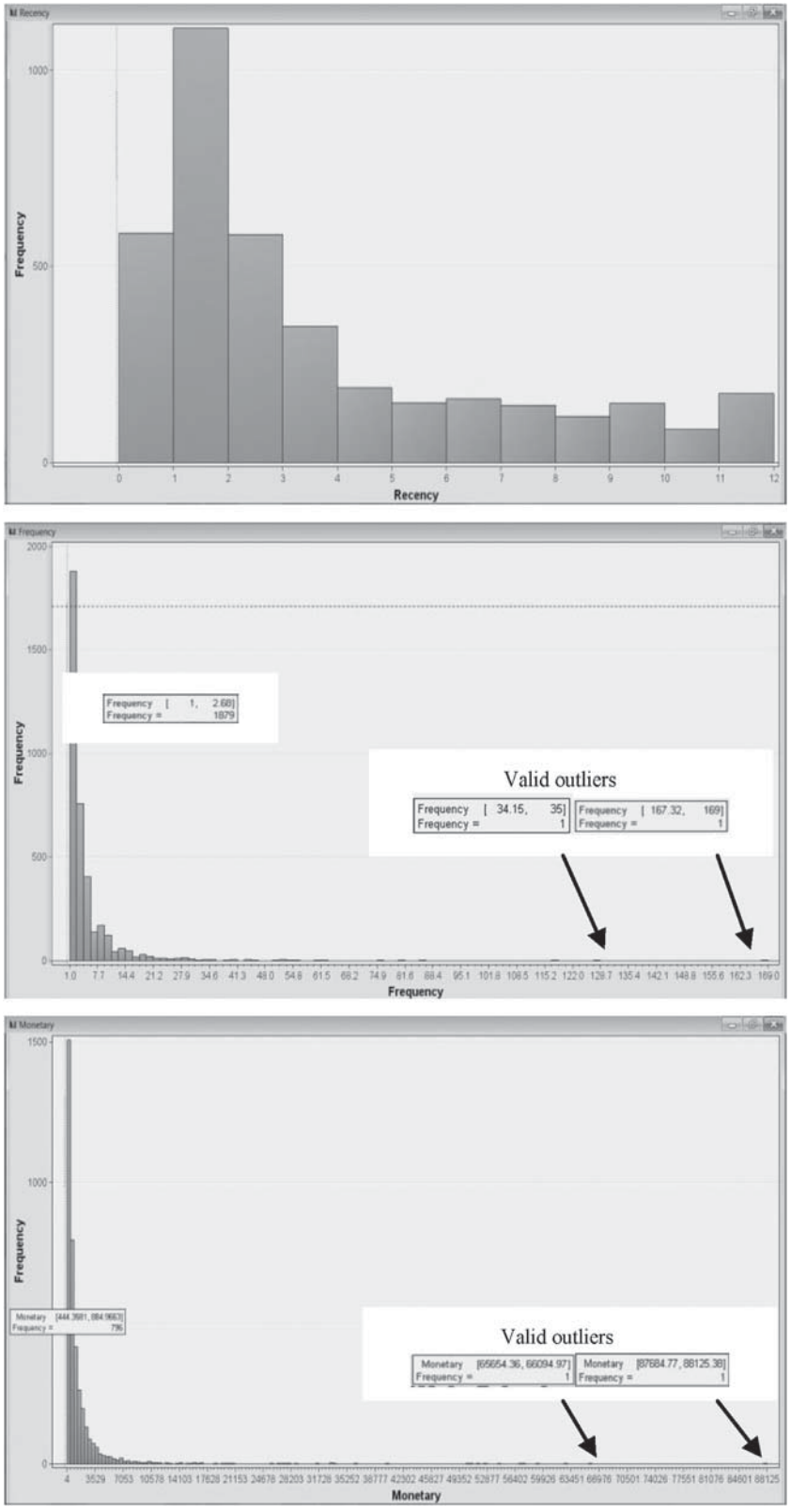

Figure 2: Distribution of the variables Recency, Frequency and Monetary.

Figure 3: Project diagram in SAS Enterprise Miner 6.2.

Table 5: Summary of the filtered target dataset (3726 instances)

Variable name	*Minimum*	*Median*	*Maximum*
Recency	0	3.2	12
Frequency	1	4.1	28
Monetary	3.75	1565.70	13110.02
First_Purchase	0	7.5	12

Table 6: Instances in each cluster

Cluster	*Frequency of cluster*	*Percentage*
1	527	14.14
2	636	17.07
3	1748	46.91
4	627	16.83
5	188	5.05

in SAS Enterprise Miner for the clustering analysis as depicted in Figure 3. There are four nodes in the diagram. In the Data Sources (Target Dataset) node, the three variables *Recency*, *Frequency* and *Monetary* were chosen as input for the clustering analysis. The Filter node was set to exclude from the analysis any instances having a rare value for any variables involved, and the minimum cutoff value for rare values was set to 1 per cent of the total number of instances under consideration. For example, out of the total 3799 instances, there was only one instance taking a monetary value of more than £87 684, and therefore, that instance was extended from the analysis. Overall there were totally 73 instances were excluded by the Filter node, and the summary of the resultant filtered target dataset is given in Table 5. In the Cluster node, the standard range transformation for normalization was used with the number of clusters specified as 3, 4 and 5, respectively, and finally, the Segment Profile node was utilized to assists to interpret each cluster found.

The clustering and segment results with five clusters are shown in Tables 6 and 7, and the distribution of the instances within each cluster is detailed in Figures 4 and 5. This segmentation by five clusters seems to have a clearer interpretation of the target dataset than the ones by three and four clusters.

Table 7: Statistics of each cluster

	Minimum	*Median*	*Maximum*
Cluster 1			
Recency	8	9.8	12
Frequency	1	1.3	4
Monetary	3.75	361.20	7741.47
First_Purchase	8	11.1	12
Cluster 2			
Recency	4	5.4	7
Frequency	1	2.3	13
Monetary	15	586.19	3906.27
First_Purchase	4	7.7	12
Cluster 3			
Recency	0	1.5	3
Frequency	1	2.6	7
Monetary	20.8	685.71	4314.72
First_Purchase	0	5.3	12
Cluster 4			
Recency	0	1.0	5
Frequency	3	8.3	16
Monetary	191.17	2425.09	7330.8
First_Purchase	1	1.0	12
Cluster 5			
Recency	0	0.7	6
Frequency	3	17.7	28
Monetary	1641.48	5962.85	13110.02
First_Purchase	0	11.1	12

Understanding the clusters

Interpreting and understanding each cluster identified is crucial in generating customer-centric business intelligence.

Examining Table 7 and Figures 4 and 5, it is interesting to see that each cluster indeed contains a group of consumers that have certain distinct and intrinsic features as detailed below.

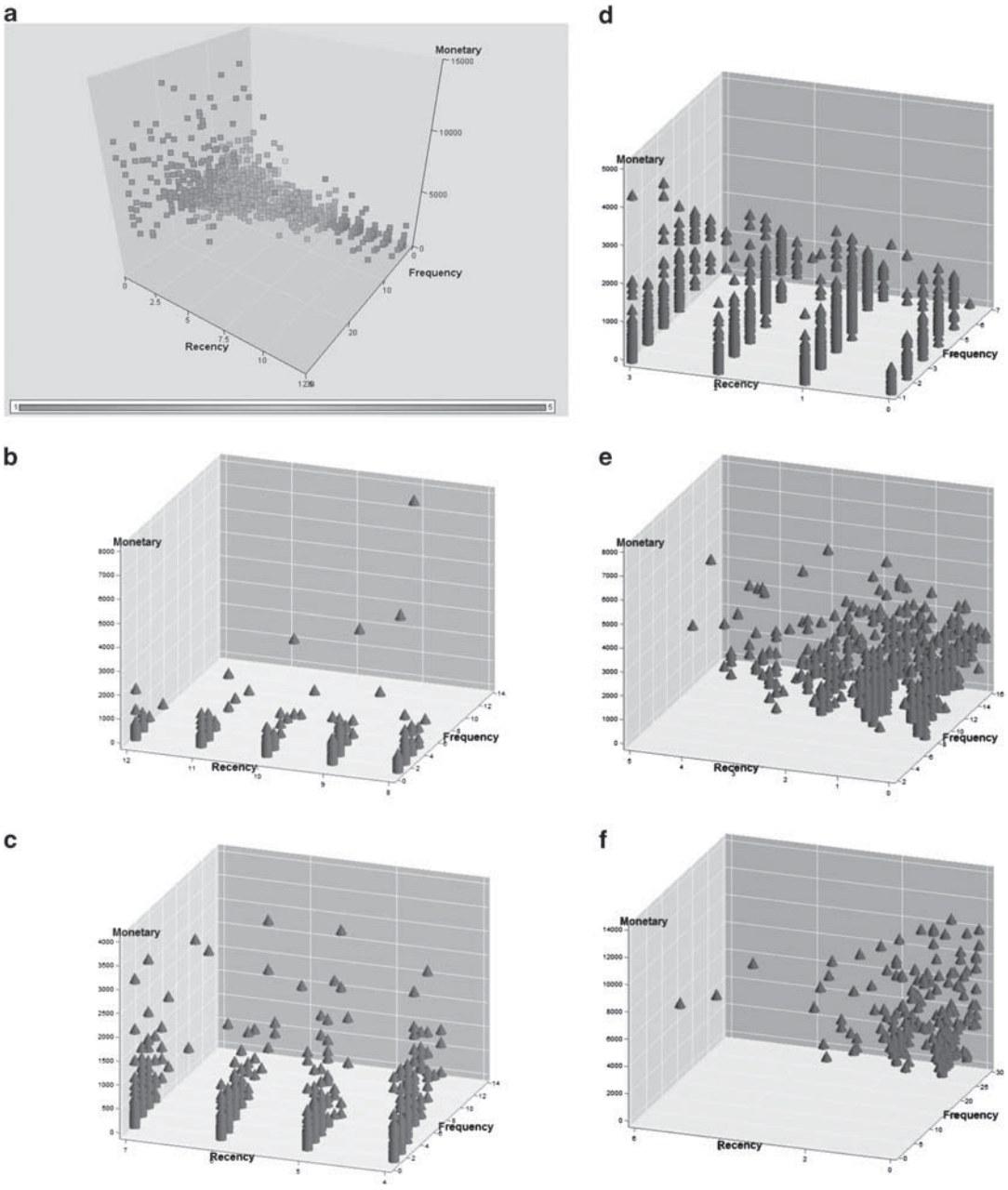

Figure 4: (**a**) Distribution of all instances coloured for different clusters. (**b**) Distribution of the instances in cluster 1. (**c**) Distribution of the instances in cluster 2. (**d**) Distribution of the instances in cluster 3. (**e**) Distribution of the instances in cluster 4. (**f**) Distribution of the instances in cluster 5.

Cluster 1 relates to some 527 consumers, composed of 14.4 per cent of the whole population. This group seems to be the least profitable group as none of the customers in this group purchased anything in the second half of the year. Even for the first half of the year, the consumers didn't shop often, and the average value of frequency was only 1.3.

Contrasted with the customers in cluster 1, the 188 customers in cluster 5 mainly started shopping with the online

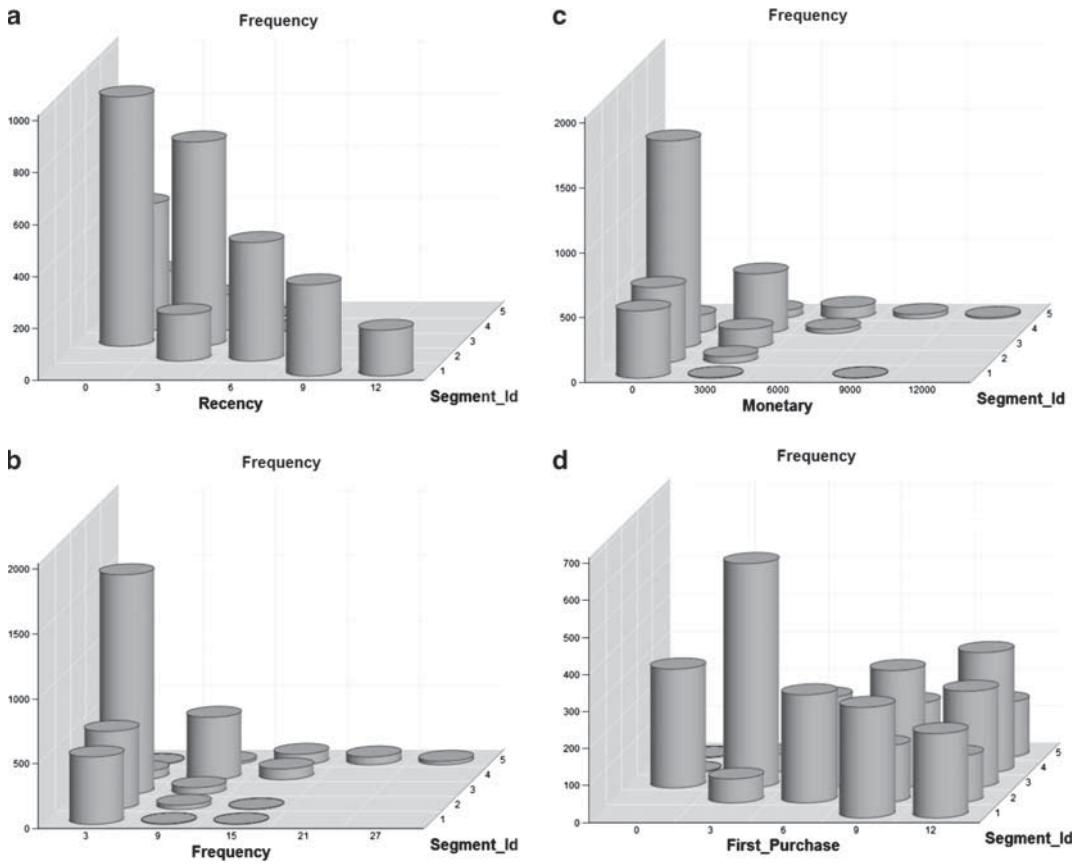

Figure 5: (**a**) Distribution of recency by cluster. (**b**) Distribution of frequency by cluster. (**c**) Distribution of monetary by cluster. (**d**) Distribution of first purchase by cluster.

retailer at the beginning of the year, and continued to the end of the year with an average value of recency 0.7. They purchased quite often and as a result, spent a quite high amount of money. This group of consumers can be categorized as very high recency, very high frequency and very high monetary with a high spending per consumer. In fact, those 188 consumers contributed 25.5 per cent of the total sales in the year. This group, although the smallest (only composed of 5.05 per cent of the whole population), seems to be the most profitable group.

Cluster 4 contains some 627 consumers with a very high value for frequency and monetary, although lower than those of cluster 5. This group seems to be the second high profit group.

There are some 459 consumers in cluster 2. Compared with clusters 4 and 5, this group of customers has a lower frequency throughout the year and a significantly smaller average value of monetary, indicating that a much smaller amount of spending per consumer. This group can be categorized as low recency, high frequency and medium monetary with a medium spending per consumer.

Cluster 3 is the largest-sized group with 1748 consumers. Consumers in this group have a reasonable value of frequency. Compared with clusters 2 and 4, this group has a lower but reasonable value of monetary as the group includes many newly registered consumers starting shopping with the retailer very recently. This group seems to have represented ordinary

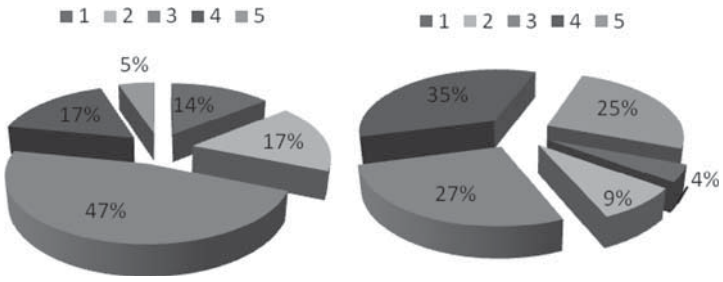

Figure 6: Customer segmentation (left) and associated sales (right) by cluster.

consumers and therefore has a certain level of uncertainty in terms of profitability. In the long-term view, some of the consumers might be potentially very highly profitable or unprofitable at all.

We use Figure 6 to summarize our analysis made so far: in the whole population of the consumers, 47 per cent of them were ordinary shoppers with reasonable spending and frequency, about 34 per cent were medium to high profit, 5 per cent were extremely highly profit, and the remaining 14 per cent were extremely low profit. About 22 per cent of the consumers contributed roughly 60 per cent of the total sales. Overall the business seems to be quite healthy in terms of profitability.

Enhancing clustering analysis using decision tree

As discussed above, cluster 3 is the most diverse cluster among the five identified clusters in the sense that it contains both newly registered and old customers as well. To refine the segmentation of the instances in this cluster, a decision tree has been used to create some nested segments internally inside the cluster, as shown in Figure 5. In other words, these nested segments form some sub-clusters inside cluster 3, and make it possible to categorize the consumers concerned into some sensible sub-categories. For example, as shown in Figure 7, the customers can be divided into such categories as frequency more than 2.5 with an average monetary value of 990.66; and frequency more than 2.5 and less than 3.5 with an average monetary value of 1056.70 and so on. Also, it is interesting to note that the relationship between frequency and monetary seems to be a monotonic linear relationship.

CUSTOMER-CENTRIC BUSINESS INTELLIGENCE AND RECOMMENDATIONS

The most valuable consumers of the business have contributed more than 60 per cent of the total sales in year 2011, whereas the least valuable ones only made up 4 per cent of the total sales. For each of these consumer groups, it is essential to further find out which products the customers in each group have purchased, which products have been purchased together most frequently and in which sequence the products have been purchased.

The business can gain a better understanding of the consumers by exploring the associations among consumer groups and the products they have purchased. The association can be examined on products/items level and on products categories level as well.

Many of the consumers of the business were organizational consumers with a high quantity of a product per transaction. Examining at which specific times (seasons), what products and which types of products they have purchased frequently will be beneficiary to the business. It will be also

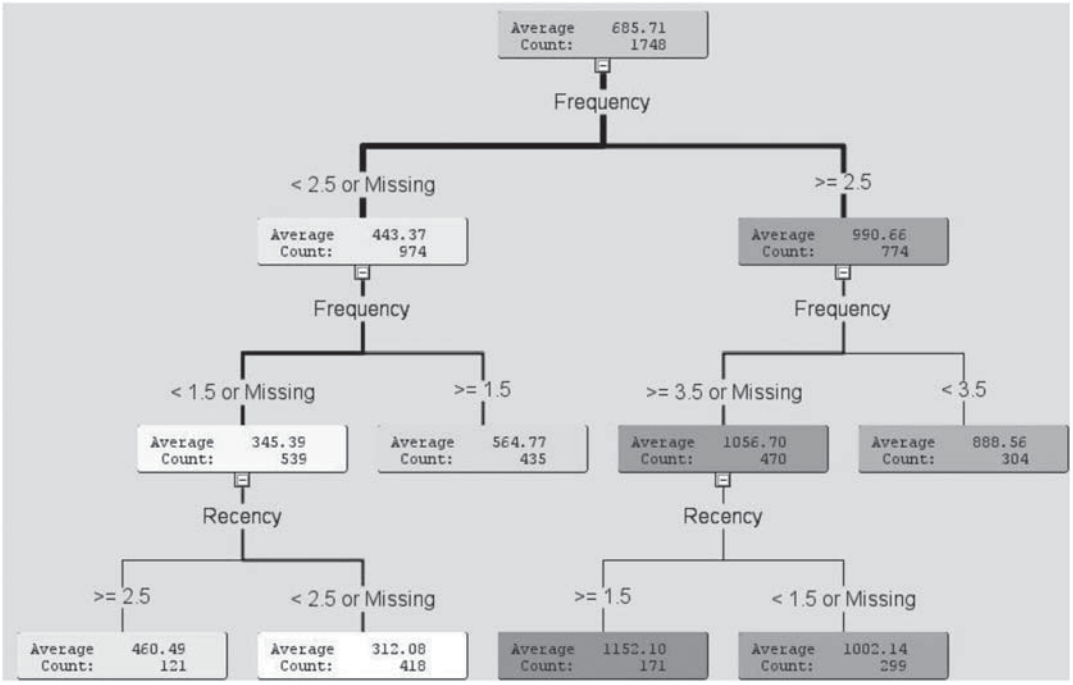

Figure 7: Refined segmentation of the instances in cluster 3 using decision tree induction.

interesting to see if there are any differences between different types of customers, that is, organizational and individual customers, in terms of their shopping patterns.

Monitoring the diversity of the most diverse customer group and predicting which customer will potentially become affiliated to the most or the least profitable group is very useful for the business in the long term. Identifying appropriate predictors or indictors for such predictions is invaluable.

Another aspect worth further investigation is to link consumer groups to geographical locations. This correlation, if exists, may help the business look into other factors, such as culture, customs, and economics, that may affect a consumer's buying intention and preferences.

CONCLUDING REMARKS

A case study has been presented in this article to demonstrate how customer-centric business intelligence for online retailers can be created by means of data mining techniques. The distinct customer groups characterized in the case study can help the business better understand its customers in terms of their profitability, and accordingly, adopt appropriate marketing strategies for different consumers.

It has been shown in this analysis that there are two steps in the whole data mining process that are very crucial and the most time-consuming: data preparation and model interpretation and evaluation.

Further research for the business includes: conducting association analysis to establish customer buying patterns with regard to which products have been purchased together frequently by which customers and which customer groups; enhancing the merchant's web site to enable a consumer's shopping activities to be captured and tracked instantaneously and accurately; and predicting each customer's lifecycle value to quantify the level of diversity of each customer.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this article.

REFERENCES

1 Interactive Media in Retail Group (IMRG). (2012) Press archive, http://www.imrg.com, accessed January 2012.
2 Kumar, V. and Reinartz, W.J. (2006) *Customer Relationship Management: A Databased Approach*, Hoboken, NJ: John Wiley & Sons.
3 Hughes, A.M. (2012) *Strategic Database Marketing 4e: The Masterplan for Starting and Managing a Profitable, Customer-based Marketing Program*, McGraw-Hill Professional, USA.
4 Davenport, T.H. (2009) Realizing the Potential of Retail Analytics: Plenty of Food for Those with the Appetite. Working Knowledge Report, Babson Executive Education.
5 Fuloria, S. (2011) *How Advanced Analytics Will Inform and Transform U.S. Retail*. Cognizant Reports, July, http://www.cognizant.com/InsightsWhitepapers/How-Advanced-Analytics-Will-Inform-and-Transform-US-Retail.pdf, accessed January 2012.
6 Collica, R.S. (2007) *CRM Segmentation and Clustering Using SAS Enterprise Miner*, Cary, NC: SAS Insititute.
7 Cerrito, P.B. (2007) *Introduction to Data Mining Using SAS Enterprise Miner*. Cary, NC: SAS Institute.
8 Sarma, K.S. (2007) *Predictive Modeling with SAS Enterprise Miner*. Cary, NC: SAS Institute.
9 Thompson, W. (2008) *Understanding Your Customer: Segmentation Techniques for Gaining Customer Insight and Predicting Risk in the Telecom Industry*. Paper 154-2008, SAS Global Forum, 16–19 March, San Antonio, TX.