Two-tiered Clustering Classification Experiments for Market Segmentation of EFTPOS Retailers

Ashishkumar Singh
Faculty of Information Technology Caulfield
Monash University
ashish.singh@monash.edu

Grace Rumantir
Faculty of Information Technology Caulfield
Monash University
grace.rumantir@monash.edu

Abstract

Almost all the papers reported in the literatures on market segmentation modeling using retail transaction data deal with finding groupings of customers. This paper proposes the application of clustering and classification techniques for finding groupings of retailers who use the Electronic Funds Transfer at Point of Sale (EFTPOS) facilities of a major bank in Australia in their businesses. The RFM (Recency, Frequency, Monetary value) analysis on each retailer is used to reduce the large data set of customer purchases through the EFTPOS network into attributes that may explain the business activities of the retailers. Preliminary results published in Singh, Rumantir and South (2014) show that groupings of retailers with distinct combinations of RFM values can be established. Encouraged by the promising business insights gained from the clustering experiments using the RFM values, in this paper, we report our findings in extracting business rules pertinent to each cluster. For this purpose, we incorporate attributes of the EFTPOS transaction data in addition to the derived RFM attributes to build a decision tree to facilitate the extraction of the business rules.

Keywords: Market Segmentation, Clustering, Decision Tree, Business Rules Extraction, RFM Analysis, EFTPOS.

1 Introduction

Electronic Funds Transfer at Point of Sale (EFTPOS) is one of the leading methods of payment in the retail industry. Payments on EFTPOS terminals are done using debit and credit cards, the most common non-cash payment methods. In 2014, over 2.4 billion EFTPOS transactions were put through over 820 thousand EFTPOS terminals in Australia. These transactions amount to over 139 billion dollars (EFTPOS Annual Report, 2015). The banking sector gains profits in providing retailer with EFTPOS machines through set up fee, periodic service fee and transaction fee on each purchase put through an EFTPOS machine. Banks also profit through the availability of interest free fund en-route to the designated retailer's bank account after being debited from the payee's account and the availability of the deposited fund into the retailer's account itself. In order for the banking sector to develop this part of the business further, it can make use of market segmentation modelling to gain better understanding of the business behaviours of the retailers using the EFTPOS facilities.

The concept of market segmentation was first introduced in Smith (1956) and defined as the “process of subdividing a market into distinct subsets of customers that behave in the same way or have similar needs. Each subset may conceivably be chosen as a market target to be reached with a distinctive marketing strategy” (Doyle 2011). The heart of any good market segmentation tool is in its ability to analyse, understand and draw good market segments based on customer purchasing behaviours. Whilst market segmentation has been extensively applied on transaction data to find insights into customer purchasing behaviours in the market, very limited work has been done to find insights into retailer business activities.

Figure 1 show the workflow followed in this paper in conducting the two-tiered experiments to find retailer business characteristics from the transactions they have generated through the EFTPOS facilities of a major bank in Australia

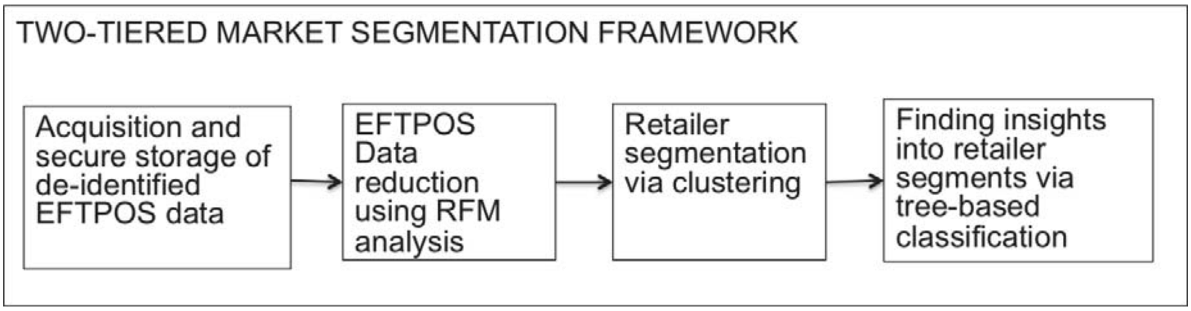
TWO-TIERED MARKET SEGMENTATION FRAMEWORK
Acquisition and secure storage of de-identified EFTPOS data
EFTPOS Data reduction using RFM analysis
Retailer segmentation via clustering
Finding insights into retailer segments via tree-based classification

Figure 1: Two Tiered Market Segmentation Framework

First, clustering techniques are employed on EFTPOS transaction data to identify segments of retailers who have certain common characteristics. The selection of attributes plays an important role in good clustering analysis. We use the RFM (Recency, Frequency, Monetary value) analysis, popular in marketing, to reduce the data set for the clustering experiments.

Second, to gain further insights into the characteristics of the retailers in each cluster in the form of common business rules, we build a decision tree using the R, F, M values and a few other EFTPOS attributes which may help explain a particular retailer, e.g. the business the retailer is in, whether or not the EFTPOS machine used belongs to the bank the retailer is a customer of, the type of card with which the transaction has been paid, etc.

The paper is organized as follows: Section 2 gives a summary of the review of the literatures on market segmentation in the retail industry in the past decade; Section 3 explains the first tier of the experiments in building the retailer segments, i.e. the data reduction/transformation using the RFM analysis, the clustering techniques used and the results of the experiments. Section 4 explains the second tier of the experiments in finding insights into the retailer segment using tree-based classification, i.e. the data pre-processing of the additional EFTPOS attributes used, the resulting classification model and the business rules extracted; Section 5 concludes the paper and explains our plans for the future.

2 Related Work

Table 1 gives the summary of the review of the literatures on related work on market segmentation in the retail industry with respect to the attribute selection techniques and clustering techniques employed. All of the papers reviewed report the use of transaction data as input for segmentation experiments on retail customers. Only one paper, i.e. Bizhani & Tarokh (2011), reports segmentation experiments on EFTPOS retailers. This is the only work on EFTPOS data we have found in the literatures. The work reported in Bizhani & Tarokh (2011) is on a much smaller data set (30,524 EFTPOS terminals with 1,030,120 total transactions) and considers individual EFTPOS machines as “retailers”. The data set we have been acquiring for our work consists of 77.5 million transaction records of over 1 million unique retailers (each retailer may use multiple EFTPOS terminals) and each transaction instance has around 60 variables. Hence, based on the volume of the EFTPOS data set used, this project can be categorised as a Big Data project. Our experience in acquiring, secured-storing and processing the commercial in confidence EFTPOS data has been reported in Singh, Rumantir, South & Bethwaite (2014).

The second and third columns of Table 1 are for the two attribute selection techniques identified in the literatures as the most popular for market segmentation in the retail industry, i.e. socio-demographic analysis and RFM analysis. Kim et al. (2005) and Lee & Park (2005) report good results on customer segmentation modelling using socio-demographic characteristics of customer households (e.g. average size of households, average age of residents, proportions of residents with different marital status, etc). Previous research suggests that customer purchase history is far better and more accurate predictors of future purchase behaviour when compared to their demographic characteristics (Gupta et al., 2006). This finding is in line with our finding as summarised in Table 1 that RFM analysis on customer transactions is more widely used as attribute selection technique in market segmentation in the retail industry.

Related Work	**Attribute Selection Techniques**		**Clustering Techniques**		
	Socio-demographic Analysis	**RFM Analysis**	**K-means**	**Hierarchical clustering**	**Other**
(Chen et al., 2012) (Lefait & Kechadi, 2010)		X	X		
(Hsieh 2004)		X			X Neural Network
(Bizhani & Tarokh, 2011)		X	X		X Unsupervised Learning Vector Quantization
(Chen et al., 2012)		X	X		
(Olson et al., 2009)		X			X
(Namvar et al., 2010)	X	X	X		
(Kim et al., 2005) (Lee & Park, 2005)	X				X Neural Networks
(Dennis et al. 2003)	X				X
(Ho et al. 2012)			X Genetic Algorithms		
(Salvador & Chan 2004)			X	X	
(D. Gaur & S. Gaur 2013)			X	X	X
(Zakrzewska & Murlewski 2005)			X	X	X Density based Clustering
(Alam et al. 2010) (Yoon et al. 2013)				X	
(Li et al. 2009)				X Chameleon	
(Suib & Deris 2008)				X Hierarchical Pattern Based Clustering	

Table 1. Summary of the review of the literatures on market segmentation in the retail industry in the past decade

Clustering is used in market segmentation because the number of possible segments in the industry is usually unknown. Chen, Sain, & Guo (2012) use K-Means clustering algorithm to segment customers of an online retail business. They pre-process the data set using the RFM analysis and then use SAS Enterprise Miner to perform the clustering analysis. Their study concludes that only 5% of the total number of customers has contributed to 25% of the total sales. Li, Wang, & Xu (2009) develop a two-stage clustering algorithm for customer segmentation of the customers of Chinese Petroleum Corp. Their research identifies a cluster of the most valuable customers to the corporation as the customers not only purchase petroleum frequently but also spend a lot of money on each purchase.

With respect to clustering algorithms used, K-means and hierarchical clustering are found to be the most popular techniques for market segmentation in the retail industry. Most of the papers use either Decision Tree or Association Rules Analysis to generate business rules pertinent in the clusters.

3 Clustering Experiments

For this project, EFTPOS transaction data are being collected from one of the four major banks in Australia. This paper reports on the preliminary stage of our project where we use data from a total period of 18 days, starting from 19 September 2013 to 7 October 2013. Each transaction record has 55 attributes. The 18 daily data files contain approximately 77.5 million transaction records from over 1 million unique retailers. This high volume of data makes even the basis operations, such as calculating the total monetary amount of each retailer from the data set, very time consuming and resource intensive. Strategies to overcome this Big Data problem is reported in Singh, Rumantir, South & Bethwaite (2014).

3.1 RFM Analysis

The RFM (Recency, Frequency, Monetary value) analysis was proposed in Hughes (2006). This paper proposes the use of these three attributes to group retailers exhibiting similar business activities:

Recency - the Recency value of a retailer is the time interval between a global datum and his/her latest transaction. A retailer with a smaller Recency value is seen to be more current in his/her business activities than a retailer with a bigger Recency value. The midnight after the day of the last transaction in the data set (i.e. the midnight of 8 October 2013 (00:00:000000 in HH:MM:SSSSSS format) is chosen as the global datum. Hence, the Recency value of each retailer is calculated from the date of his/her latest transaction up to midnight on 8 October 2013.

Frequency - the total number of transactions put through all of the EFTPOS terminals belonging to a retailer forms the Frequency value of the retailer.

Monetary value - the total amount of transactions put through all of the EFTPOS terminals belonging to a retailer forms the Monetary value of the retailer.

Figure 2 shows the bar chart of the distribution of the Recency, Frequency and Monetary values of all of the retailers in the data set. The bar chart shows that the Frequency and Monetary values are skewed to the lower end of the spectrum.

The bar chart in Figure 2 is open ended at the top end because there are small numbers of very large values of the three attributes in the data set. This is shown more clearly in the bar charts in Figure 3 where the horizontal axis of each bar chart is intentionally “squeezed” in the middle to show these extreme values and also because there are very few data in this range (no data for Frequency and Monetary values).

Figure 3 also shows the correlations amongst the three attributes. There is positive correlation between Frequency and Monetary values and no correlation between Recency with the other two attributes. This implies that retailers that generate a lot of transactions tend to accumulate large Monetary values within the observation time period.

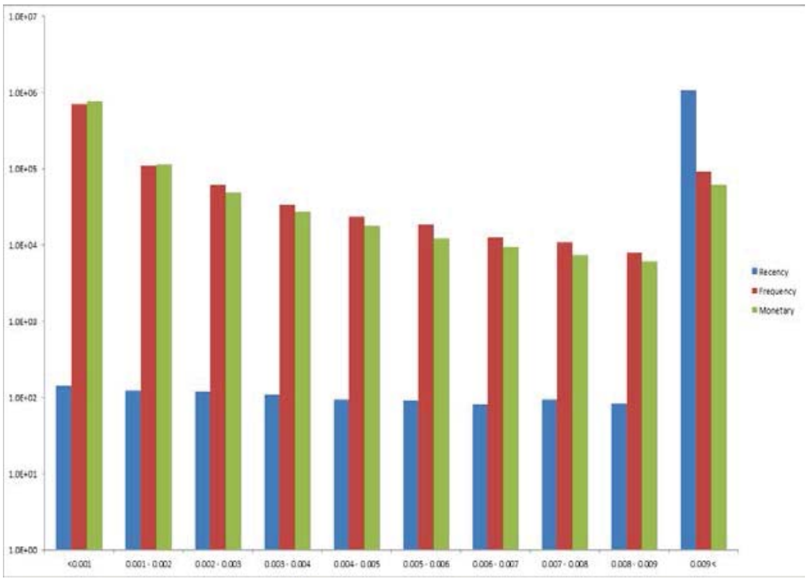

Figure 2. The distribution of Recency, Frequency and Monetary values of all retailers

3.2 Clustering Techniques

Being the most popular methods for market segmentation in the retail industry found in the literatures, we use K-means and Agglomerative Hierarchical Clustering (AHC) in this preliminary work. Clustering analysis on a large data set is time consuming and processor intensive. For this work, parallelisation in the form of multi-threading using a cluster of Intel Xeon and AMD Opteron CPUs of various clock speeds with 16 cores, 32 GB RAM and 500GB hard disk size is employed. Our experience in acquiring, secured-storing and processing the commercial in confidence EFTPOS data is reported in Singh, Rumantir, South & Bethwaite (2014).

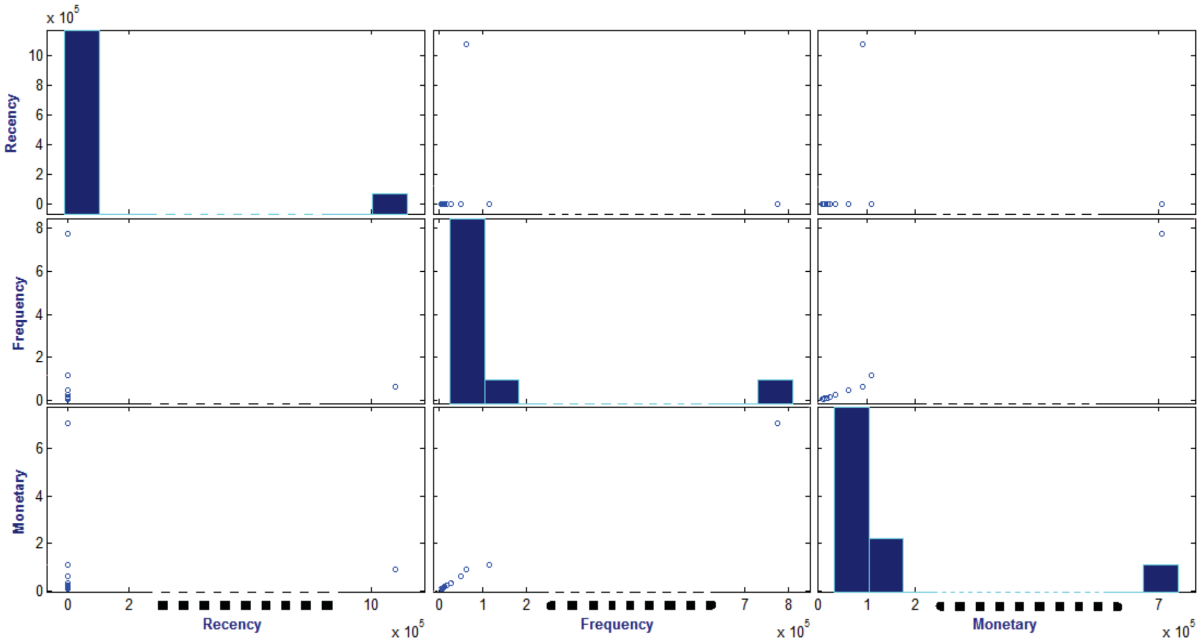

Figure 3. The correlations amongst the three attributes. The middle of each of the x-axis are deliberately truncated to show the small number of extreme values at the upper end, made possible as there are no data Frequency and Monetary values and very few Recency values in the middle range

For the K-means clustering, the calculation of the Euclidean Distance between each data point to the centroid of a cluster is done in parallel over subsets of the data set. For the AHC, the calculation of the Ward Minimum Variance to evaluate the inter-cluster distance measure (starting with clusters with 1 member data point) and the creation of the clusters through agglomerative process are done on 60% of the data set with the remaining 40% are allocated to the clusters based on Euclidean distance.

In this project, we create clustering models with 2 to 25 clusters using both K-means and AHC. Both sets of clustering models are then tested based on Dunn's Index (Dunn† 1974) and Davis-Bouldin Index (Davies & Bouldin 1979) to find the optimum number of clusters which result in high intra-cluster similarity and high inter-cluster dissimilarity.

Dunns Index is given as:

$$D = \min_{1 \leq i \leq c} \left\{ \min_{1 \leq j \leq c, j \neq i} \left\{ \frac{\delta\,(C_i\,,\,C_j)}{\max\limits_{1 \leq k \leq c} \{\Delta(C_k)\}} \right\} \right\}$$

where: $\delta\,(C_i\,,\,C_j)$ is the inter-cluster distance between clusters c_i and c_j; $\Delta(C_k)$ is the intra-cluster distance of cluster c_k; c is the number of clusters.

Also Davis-Bouldin Index is given as:

$$DB = \frac{1}{c} \sum\nolimits_{i=1}^{c} \max_{i \neq j} \left\{ \frac{\Delta\,(C_i) + \Delta\,(C_j)}{\delta\,(C_i\,,\,C_j)} \right\}$$

where: $\delta\,(C_i\,,\,C_j)$ is the inter-cluster distance between two clusters c_i and c_j; $\Delta(C_i)$ and $\Delta(C_j)$ are the intra-cluster distances of clusters c_i and c_k respectively; c is the number of clusters.

Two inter-cluster distance measures and two intra-cluster distance measures are used to evaluate these two indices as outlined in Table 2.

Intercuster distance measure	Average linkage distance measure	Centroid linkage distance measure
	$\Delta\,(C) = \frac{1}{\lvert C\rvert \times (\lvert C\rvert - 1)} \sum_{x \in C,\ y \in C,\ x \neq y} d\ (x\,,y)$	$\Delta\,(C) = 2\,(\frac{\sum_{x \in C} d\,(x\,,\ \bar{C})}{\lvert C\rvert})$ where $\bar{C} = \frac{1}{\lvert C\rvert} \sum_{x \in C} x$
Intracluster distance measure	Average diameter distance measure	Centroid diameter distance measure
	$\delta\,(C_X\,,C_Y) = \frac{1}{\lvert C_X\rvert\lvert C_Y\rvert} \sum_{x \in C_X,\ y \in C_Y} d(x\,,y)$	$\delta\,(C_X\,,C_Y) = d\,(\frac{1}{\lvert C_X\rvert} \sum_{x \in C_X} x\,,\ \frac{1}{\lvert C_Y\rvert} \sum_{x \in C_Y} y)$

Table 2. Two pairs of inter-cluster and intra-cluster distance measures to be used in Dunns and Davies Bouldin Indices.

3.3 Clustering Results and Analysis

The Dunns and Davies–Bouldin indices for each of the 24 clustering models (with 2 to 25 clusters respectively) created using each of the two clustering techniques are plotted in Figure 4 and Figure 5 respectively. Comparing the results between the K-means and AHC models, we can see that the AHC models with 16 and 17 clusters have higher Dunn's indices than any of the K-means models. The K-means models with 13 clusters or more tend have higher Davies–Bouldin indices than the AHC models with the same number of clusters. It is then decided to analyse the AHC model with 19 clusters with a view that manual merging of clusters might be warranted post-analysis.

Table 3 shows the AHC clusters, each with the centroid represented by the average RFM values. Each of the R, F, M values in the data set is divided into 3 quantiles, with labels 1 to 3. Based on the centroid values, each cluster is assigned a 3 digit label which values depend on which range each of its R, F, M levels falls in.

Low level in Recency suggests the retailer has recent transactions. Hence, low level in Recency and very high levels in both Frequency and Monetary value suggest the feature of a retailer with active, successful business. In Table 3, these retailers (labelled as Active Profitable Retailers) can be found in Clusters 1, 2, 18 (constitute 34.23% of the total retailers). The bank may wish to provide incentives for these retailers to maintain their excellent business performance.

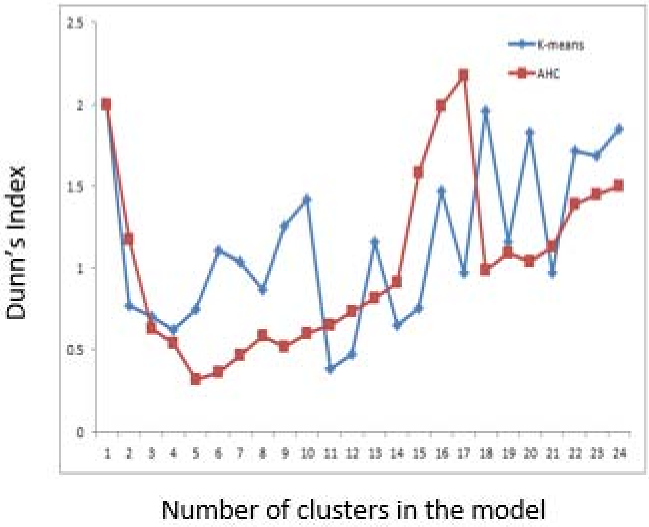

Figure 4. A plot of the Dunns index of each model created using K-means technique and the Dunns index of each model created using Agglomerative Hierarchical Clustering technique

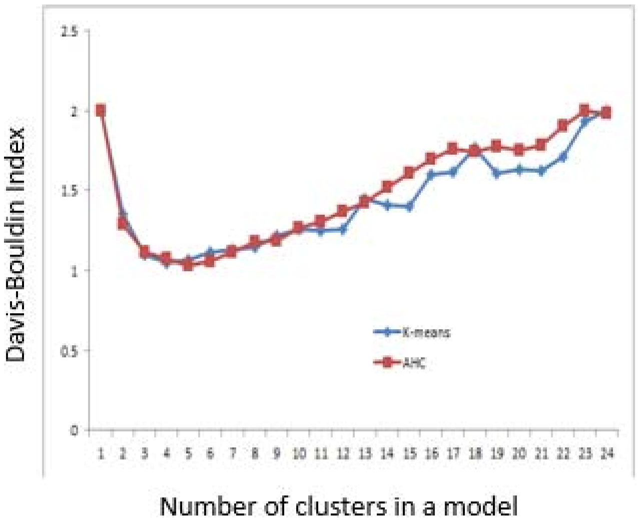

Figure 5. A plot of the Davis-Bouldin index of each model created using K-means technique and the Dunns index of each model created using Agglomerative Hierarchical Clustering technique

There are retailers with high levels of Frequency and Monetary values but with moderately high Recency level (in Clusters 15, 16, and 19, which constitute 30.26% of the total retailers). These retailers (labelled as Recently Less Active Retailers) are very valuable to the bank as they have generated a lot of transactions and had a high revenue level despite having been less active lately. However they are at risk of becoming even less active which will inadvertently lower their Frequency and Monetary levels. The bank needs to take immediate remedial actions to ensure that this risk can be alleviated.

Retailers characterized by moderately high Recency level with medium levels of Frequency and Monetary values (labelled as Used to be Active Profitable Retailers) are seen in Clusters 3, 4, 5, 6, 11, 12, 13, 14 and 17 which constitute 27.63% of the total retailers. These retailers have businesses with just moderate level of activities obviously as a result of having been inactive for some time. The bank may want to engage these retailers into more active participation in generating EFTPOS transactions as this will consequently boost the levels of their Frequency and Monetary values.

Cluster #	Amount of data (%)	Average Recency	Average Frequency	Average Monetary Value	R,F,M labels (1 to 3 quantiles)
1	**10.30**	**2.368704**	**0.004604**	**0.003177**	**133**
2	**15.27**	**1.018574**	**0.038917**	**0.017611**	**123**
3	2.41	67.028300	0.000117	0.000333	322
4	2.30	71.897224	0.000105	0.000313	322
5	1.72	47.020206	0.000343	0.000331	322
6	3.20	42.105650	0.000293	0.000681	322
7	1.31	82.475670	0.000085	0.000163	312
8	2.14	76.985930	0.000071	0.000287	312
9	2.11	97.306885	0.000025	0.000220	312
10	2.32	92.311350	0.000067	0.000303	312
11	3.25	57.261173	0.000191	0.000473	322
12	2.79	62.190346	0.000169	0.000406	322
13	5.12	27.091051	0.000393	0.000818	323
14	4.46	32.134422	0.000318	0.000703	322
15	7.36	21.853570	0.000626	0.001149	223
16	11.78	17.150494	0.001816	0.002548	233
17	2.39	52.247010	0.000147	0.000263	322
18	**8.66**	**6.925114**	**0.011165**	**0.004900**	**133**
19	11.12	12.214928	0.002484	0.002045	233

Table 3. Clustering Results

Retailers characterized by very high Recency level with low level of Frequency value and medium level of Monetary value are seen in Clusters 7, 8, 9 and 10. These retailers have been inactive in the EFTPOS network for a long period of time and/or have low level business activities overall. They constitute 7.88% of the EFTPOS market for the bank. The bank can see this type of retailers as growth area in the EFTPOS business sector, hence can be labelled as Growth Area Retailers. It however needs to determine if it is cost effective to launch aggressive marketing strategy to help improve the performance of these retailers. If it is decided that no aggressive campaigning is to be done for these retailers, due to the size of this market, some kind of marketing campaign will nevertheless still be required to maintain them. Figure 6 shows the aggregate breakdown of the retailers.

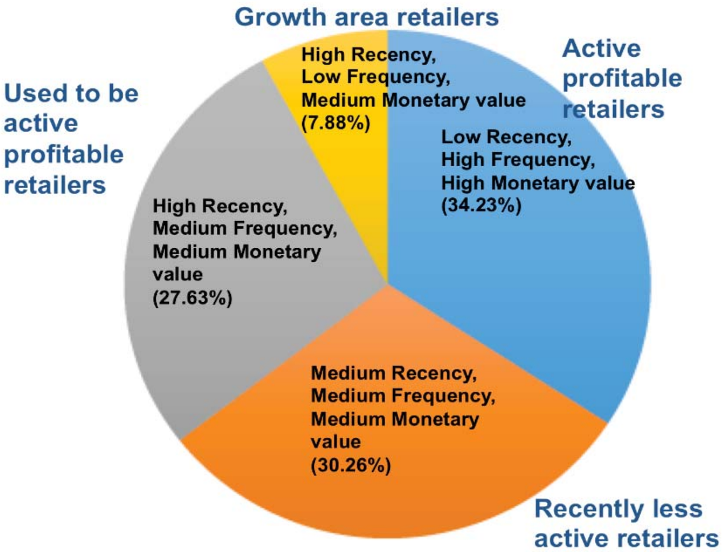

Figure 6. Retailer categories: aggregate breakdown of the 19 clusters of retailers shown in Table 3

4 Decision Tree Experiments

To find further insights into the 19 retailer clusters, a decision tree is built using the R, F and M values together with other transaction attributes that may have information to better understand the characteristics of the retailer who has generated the respective transactions. The process of building the decision tree will determine if an attribute used in the experiments is indeed a relevant characteristic of a retailer.

4.1 Data Preprocessing

In addition to the R, F and M values, Table 4 shows the attributes in the EFTPOS transaction data that are used in the Decision Tree experiments. Data pre-processing in the form of reducing of the number of possible values an attribute can have is done for most of the attributes. The link between the Retailer's Financial Institution ID (FIID) and the Customer's FIID can be explored further. Since we are only interested in the transactions that have to do with the retailers of the bank under study (we call it TheBank), we can just use binary values for these attributes. Also, we observe that there are the following 4 possible combinations of values in the transactions. This leads us to replace both attributes with a new combined attribute having 4 possible pairs of binary values:

TheBank's EFTPOS machine? = YES and TheBank's card? = YES → 11

TheBank's EFTPOS machine? = YES and TheBank's card? = NO → 10

TheBank's EFTPOS machine? = NO and TheBank's card? = YES → 01

TheBank's EFTPOS machine? = NO and TheBank's card? = NO → 00

No	Name	Description	Original Values	Pre-processed Values	End up in decision tree?
1	Standard Industrial Classification (SIC) Code	A four-digit code to identify the industry the retailer is in	A four-digit code representing division, sub-division, group and class	1 digit code with a total of 11 divisions	✓
2	Retailer's Financial Institution ID (FIID)	The FIID of the institution with which the retailer is associated – The owner of the EFTPOS Machine at Retailer	The name of a bank	TheBank or other bank (Yes/No)	✓
3	Customer's Financial Institution ID (FIID)	The FIID of the institution that issued the card	The name of a bank	TheBank or other bank (Yes/No)	✗
4	Transaction Code	A code identifying the type of transaction	18 types	3 new types: Purchase Related, Cash Advanced, Enquiries	✗
5	Card Type	A code identifying the card type associated with the transaction	3 types	debit, credit, other	✗
6	Account Type	A code identifying the type of account associated with the transaction.	4 types	DDA, savings, credit, none	✗

Table 4. EFTPOS attributes used for Decision Tree experiments

4.2 Decision Tree Induction

As illustrated in Figure 7, a decision tree starts with a root node and ends with leaf nodes. In our project, the root node is the attribute that is the most useful in identifying which class a retailer belongs to. The leaf nodes represent the classes. In the middle are the child nodes, each child node is an attribute with branches representing the possible values of the attribute. Business rules can be extracted by traversing down the decision tree from the root node to each of the leave nodes. In this project, the 19 clusters a retailer can belong to are the 19 classes of the decision tree.

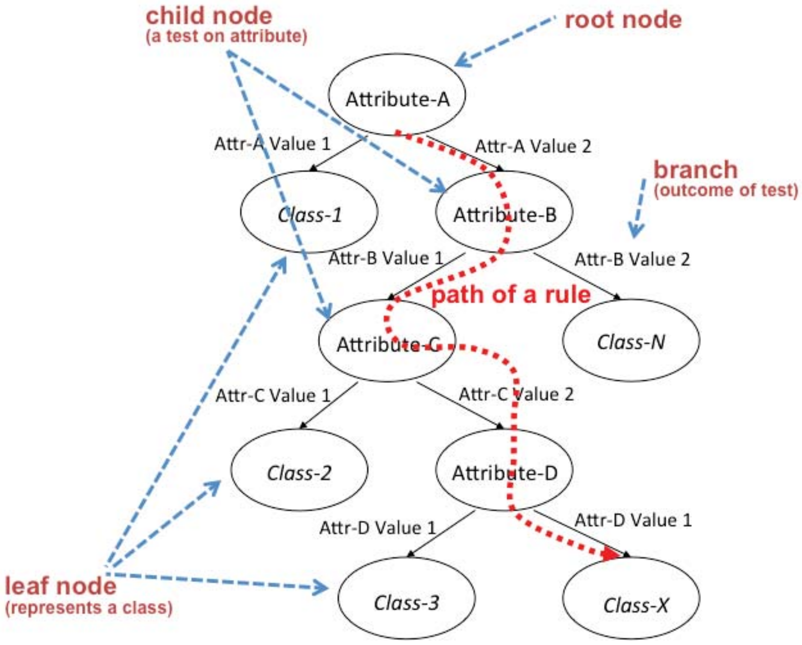

Figure 7. A generic decision tree

To build the decision tree we need to determine which attribute to place at the root node, and the subsequent child nodes. The concept of Information Gain is used to make this choice at each node. Information Gain is the difference between the amount of information that is needed to make a correct classification before a decision/split at a node is made, and the amount of information required after the split. If the amount of information required after the split is lower, then the split is said to have decreased the disorder in the original data. The attribute that yields the most Information Gain will be selected to make the split at a node.

$$InfoGain = InfoRequired_{beforeSplit} - InfoRequired_{afterSplit}$$

We use entropy, the most common measure of the amount of information at a node:

$$I = \sum_i -(p_i \log_2 p_i) \qquad \text{(Equation 1)}$$

where: p_i is the fraction of the instances classified as class *i*.

In this paper, once the decision tree has been created, pruning is done by removing child nodes which (1) have been selected based on very small amount of Information Gain and (2) lead to leaf nodes with very small number of instances. This results in only two of the attributes in Table 4, i.e. SIC code and Retailer's FIID (i.e. if the EFTPOS machine belongs to TheBank) ending up in the final decision tree. This means the type of card used in a transaction is not significant for the purpose of retailer classification. Majority voting is the simplest most common way of labelling a cluster/class. Majority voting is employed for classifying a leaf node where not all instances belong to the same class.

The decision tree is built using 80% of the total number of retailers and is tested using the remaining 20%. The performance of the decision tree is evaluated by calculating the classification accuracy of the retailers into their respective clusters:

$$Accuracy = \frac{NumberOfCorrectlyClassifiedRetailers}{NumberOfRetailers} \qquad \text{(Equation 2)}$$

The final decision tree for the retailer classification shown in Figure 8 has 72% accuracy on test data.

4.3 Business Rules Extraction and Analysis

The decision tree shown in Figure 8 has Recency as the root node. This means the results of the entropy calculation used to build the decision tree is consistent with the findings in marketing and advertising literatures that out of the three values in RFM analysis, Recency is the most important factor in predicting the likelihood of a customer performing a repeat business based on past purchasing behaviour.

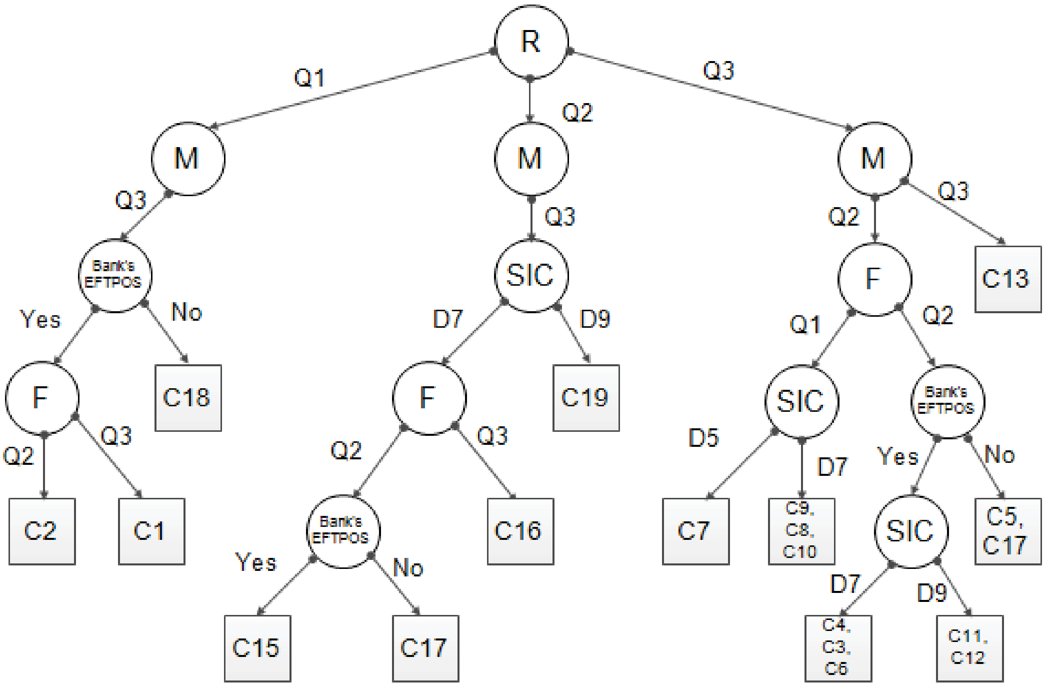

Figure 8. Decision tree for the retailer classification into segments using the R, F, M values and the additional EFTPOS attributes

In Figure 8, R, F and M represent Recency, Frequency and Monetary values respectively. Q1, Q2 and Q3 are the three quantiles of the RFM values. C1 to C19 represents clusters 1 to 19. D5 (Transportation, Communication, Electric, Gas and Sanitary Services), D7 (Retail Trade) and D9 (Services) are the industry divisions of the SIC codes.

In Figure 8 we see that Monetary value and Frequency are high up in the tree as expected. This means the additional EFTPOS attributes, i.e. whether or not the EFTPOS machine belongs to TheBank and SIC code, serve as additional explanations of the retailer clusters which have been built with just using the three attributes, Recency, Frequency and Monetary value. More specifically, by using the retailer categories given in the pie chart in Figure 6, we can see that the left hand side of the decision tree in Figure 8 lead to the classification retailers into the Active Profitable Retailers category (Clusters 1, 2, 18). In addition to the RFM characteristics that have been identified from the clustering experiments, the decision tree shows that retailers in Clusters 1 and 2 use TheBank's EFTPOS machines while retailers in Cluster 18 doesn't, they use other bank's EFTPOS machines. The bank may see this as a business opportunity to offer the retailers in Cluster 18 to use TheBank's EFTPOS machines.

With the Recently Less Active Retailers category (Clusters 15, 16 and 19), the decision tree reveals that the retailers in Cluster 19 operate in the Services industry division and those in both Clusters 15 and 16 operate in the Retail Trade industry division. Also since it is revealed that retailers in Cluster 15 use TheBank's EFTPOS machines. This means, if the bank would like to launch a remedial action in a cost effective manner, it can start with retailers in Cluster 15.

With the Used to be Active Profitable Retailers, the decision tree reveals that the retailers in Cluster 17 are in the Retail Trade division and do not use TheBank's EFTPOS machines. The decision tree does not provide additional insights in to the other retailer clusters in this category (i.e. Clusters 3, 4, 5, 6, 11 and 12). Further investigations, e.g. in the form of using more specific SIC codes and finding other more relevant attributes, might need to be done.

With the Growth Area Retailers category, the decision tree reveals that the retailers in Cluster 7 are in Transportation, Communication, Electric, Gas and Sanitary Services industry division. This may indicate that TheBank's EFTPOS market share in this industry division may be shrinking. Table 5 gives the summary of the rules extracted from the decision tree. Each rule is evaluated using the Lift metric:

$$Lift_i = \frac{ConfidenceOfRule_i}{SupportOfCluster_i} \qquad \text{(Equation 3)}$$

where:

$$ConfidenceOfRule_i = \frac{NumberOfDecisionTreeClassification_i}{NumberOfTotalClassification_i}$$

$$SupportOfCluster_i = \frac{NumberOf\,\mathrm{Re}\,tailersInCluster_i}{TotalNumberOfRetailers}$$

It shows that all the rules have Lift values much more than 1 which indicates that the rules are significant.

Retailer Category	Rule	Lift
Active Profitable Retailers	(R=Q1 && M=Q3 && Bank's EFTPOS=Yes && F=Q3) => **C1**	6.94
	(R=Q1 && M=Q3 && Bank's EFTPOS=Yes && F=Q2) => **C2**	4.64
	(R=Q1 && M=Q3 && Bank's EFTPOS=No) => **C18**	8.44
Recently Less Active Retailers	(R=Q2 && M=Q3 && SIC=D7 && F=Q2 && Bank's EFTPOS=Yes) => **C15**	9.60
	(R=Q2 && M=Q3 && SIC=D7 && F=Q3) => **C16**	5.85
	(R=Q2 && M=Q3 && SIC=D9) => **C19**	5.89
Used to be Active Profitable Retailer	(R=Q3 && M=Q2 && F=Q1 && SIC=D7) => **(C3, C4, C6)**	Needs Further Investigation
	(R=Q3 && M=Q2 && F=Q2 && Bank's EFTPOS=No) => **C5**	
	(R=Q3 && M=Q2 && F=Q2 && Bank's EFTPOS=Yes && SIC=D9) => **(C11, C12)**	
	(R=Q3 && M=Q3) => **C13**	15.84
	(R=Q2 && M=Q3 && SIC=D7 && F=Q2 && Bank's EFTPOS=No) => **C17**	15.88
Growth areas	(R=Q3 && M=Q2 && F=Q1 && SIC=D5) => **C7**	16.64
	(R=Q3 && M=Q2 && F=Q1 && SIC=D7) => **(C8, C9, C10)**	Needs Further Investigation

Table 5. Business rules extracted from the decision tree in Figure 8

5 Conclusions and Future Work

This paper proposes the use of clustering and classification techniques to analyse retailers on an EFTPOS network. The clustering experiments group the retailers based on the similarities in their business activities as characterized by how recent their business activities are, how frequent they conduct their business on the EFTPOS network and how much money their business activities have generated over a period of time. The preliminary results show that there are distinct combinations of RFM values of retailers in the clusters that may give the bank indications of the different marketing strategies that can be applied to each of the retailer types.

To further analyse each cluster, we build a decision tree to find out more on the characteristics of the business and the background of the retailers. Out of the six additional EFTPOS attributes used to build the decision tree, two attributes namely the Standard Industrial Classification (SIC) code of the retailers and whether or not the retailer uses the EFTPOS machine belonging to the bank under study, have been calculated to have significant influences to the classification of a retailer into a cluster, hence are included in the final decision tree. These attributes provide further explanations into the characteristics of the retailers in their respective clusters.

The next step of this project will be broken down into 3 categories. First, observing if there are latent variables in the data set that may influence the variations in the volume of transactions in different days and possibly different periods in a day. Second, experimenting other clustering techniques to see if better quality clusters can be formed. Third, building improved classification or causal models to find explanatory rules on the characteristics of each cluster by using more specific SIC code and adding exogenous variables like socio-demographic, advertising, and social media data.

The data set used in this preliminary work is just from the first 18 days of our EFTPOS data extraction. We have since collected a few months of EFTPOS transaction data that will allow us to conduct more extensive Big Data analysis with consequently more convincing results. This will also open up new research avenues into the kinds of suitable Big Data computing techniques for market segmentation projects involving MapReduce/Hadoop system that we have put in place for this project.

Acknowledgments

The authors wish to thank Annie South, Jon Scheele and Jo Spencer for their help in data acquisition and Blair Bethwaite for his help in setting up the secure computing environment to enable the execution of this commercial in confidence Big Data project.

References

Alam, S. et al., 2010. Particle swarm optimization based hierarchical agglomerative clustering. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. pp. 64–68.

Bizhani, M. & Tarokh, M.J., 2011. Behavioral rules of bank's point-of-sale for segments description and scoring prediction. *Int. J. Industrial Eng. Comput*, 2, pp.337–350.

Chen, D., Sain, S.L. & Guo, K., 2012. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing \& Customer Strategy Management*, 19(3), pp.197–208.

Chen, Y.-S. et al., 2012. Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment. *Computers in Biology and Medicine*, 42(2), pp.213–221.

Davies, D.L. & Bouldin, D.W., 1979. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2), pp.224–227.

Dennis, C. et al., 2003. Market segmentation and customer knowledge for shopping centers. In *Information Technology Interfaces, 2003. ITI 2003. Proceedings of the 25th International Conference on*. pp. 417–424.

Doyle, C., 2011. *A dictionary of marketing*, Oxford University Press.

Dunn†, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1), pp.95–104.

EFTPOS Annual Report, EFTPOS Australia, http://www.eftposaustralia.com.au/about/annual-reports/ (accessed July 2015)

Gaur, D. & Gaur, S., 2013. Comprehensive Analysis of Data Clustering Algorithms. In *Future Information Communication Technology and Applications*. Springer, pp. 753–762.

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N.,Sriram, S. (2006). Modeling customer lifetime value. Journal of Service Research, 9(2), 139–155.

Ho, G.T. et al., 2012. Customer grouping for better resources allocation using GA based clustering technique. *Expert Systems with Applications*, 39(2), pp.1979–1987.

Hsieh, N.-C., 2004. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications*, 27(4), pp.623–633.

Hughes, A.M., 2006. *Strategic database marketing*, McGraw-Hill.

Kim, Y. et al., 2005. Customer targeting: A neural network approach guided by genetic algorithms. *Management Science*, 51(2), pp.264–276.

Lee, J.H. & Park, S.C., 2005. Intelligent profitable customers segmentation system based on business intelligence tools. *Expert Systems with Applications*, 29(1), pp.145–152.

Lefait, G. & Kechadi, T., 2010. Customer Segmentation Architecture Based on Clustering Techniques. In *Digital Society, 2010. ICDS'10. Fourth International Conference on*. pp. 243–248.

Li, J., Wang, K. & Xu, L., 2009. Chameleon based on clustering feature tree and its application in customer segmentation. *Annals of Operations Research*, 168(1), pp.225–245.

Namvar, M., Gholamian, M.R. & KhakAbi, S., 2010. A two phase clustering method for intelligent customer segmentation. In *Intelligent Systems, Modelling and Simulation (ISMS), 2010 International Conference on*. pp. 215–219.

Olson, D.L. et al., 2009. Comparison of customer response models. *Service Business*, 3(2), pp.117–130.

Salvador, S. & Chan, P., 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*. pp. 576–584.

Singh, A., Rumantir, G., South, A., Bethwaite, B.: Clustering experiments on big transactional data for market segmentation. Proceedings of the Third ASE International Conference on Big Data Science and Computing, Beijing (2014), ACM 978-1-4503-2891-3/14/08, http://dx.doi.org/10.1145/2640087.2644161

Singh, A., Rumantir, G., South, A.: Market Segmentation of EFTPOS Retailers. Proceedings of the Twelfth Australasian Data Mining Conference, Brisbane, Conferences in Research and Practice in Information Technology, Vol. 158. Richi Nayak, Xue Li, Lin Liu, Kok-Leong Ong, Yanchang Zhao, Paul Kennedy Eds (2014) (in press)

Smith, W.R., 1956. Product differentiation and market segmentation as alternative marketing strategies. *The Journal of Marketing*, 21(1), pp.3–8.

Suib, D.S. & Deris, M.M., 2008. An efficient hierarchical clustering model for grouping web transactions. *International Journal of Business Intelligence and Data Mining*, 3(2), pp.147–157.

Yoon, S.-H. et al., 2013. A data partitioning approach for hierarchical clustering. In *Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication*. p. 72.

Zakrzewska, D. & Murlewski, J., 2005. Clustering algorithms for bank customer segmentation. In *Intelligent Systems Design and Applications, 2005. ISDA'05. Proceedings. 5th International Conference on*. pp. 197–202.