A hybrid approach for managing retail assortment by categorizing products based on consumer behavior

MSc Research Project
Data Analytics

Dhiraj Karki
x17126282

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Dhiraj Karki
Student ID:	x17126282
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Noel Cosgrave
Submission Due Date:	13/08/2018
Project Title:	A hybrid approach for managing retail assortment by categorizing products based on consumer behavior
Word Count:	4780

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	16th September 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A hybrid approach for managing retail assortment by categorizing products based on consumer behavior

Dhiraj Karki
x17126282
MSc Research Project in Data Analytics

16th September 2018

Abstract

Managing product assortment and shelfspace has been a challenge for every retailer. Retailers face the decision on what products to keep and how much quantity. Assortment and inventory management practices can have a considerable impact on the overall business of the retailer. Studies in assortment management have been limited to understanding transactional data and creating rules for making assortment decision. Hardly any product level information apart from sales is used while making these decisions. This study focusses on understanding product-customer relationship and using it as an input for managing assortment. Certain products and categories may have significant impact on customer buying behavior and therefore it is important to identify and categorize such products based on their impact on customer behavior. For this study ideal customer segments were identified using unsupervised k-means clustering. Products we clustered into different categories using fuzzy c-means clustering method. The product buying behavior of ideal customer cluster was studied to identify products which are preferred by them using association rule mining ARM. Based on their preference these products were assigned extra weights or minimum threshold. Weighted association rule mining WARM method was used to create assortment rules, which were then compared to a general assortment strategy to test whether certain category of products need to have extra weight or minimum support threshold based on their impact on customer buying behavior.

Keywords: Assortment, customer behavior, product categorization, k-means, Fuzzy C-means, association rule mining ARM, weighted association rule mining WARM.

1 Introduction

Over the past few years, there has been a tremendous expansion in the number of product categories and SKUs. New and new products with a variety of features are launched to cater to the demands of the consumers. Consumers are also aware of the new products, SKUs and their features due to word of mouth publicity, infomercials or from social media. Therefore, a buyers expectation in terms of choices increases every time while making a buying decision. Unavailability of choices or options may severely impact a person buying behavior. From a customer relationship point of view, this may result in customer churn, downgrade in value or cannibalization into other categories. This may not be an ideal scenario for any retailer. Therefore, for most retailers store assortment and inventory management has become a key decision-making process. Retailers also understand that with the right product assortment method they can influence a customer buying behavior which would ultimately lead to a better customer relationship. At present most assortment practice involves understanding relationships between products sold and then using it to manage assortment. Most assortment methods fail to incorporate a customer behavioral approach while making assortment decisions.

For e.g. In a large supermarket store there are number of different SKUs belonging to different product categories. Wine is a type of product category and has a number of SKUs under it. Now within this category there are different variety of wines (brand, red, white etc.). As wine would not be most frequently brought item in the supermarket, wine products or category may suffer from lower support level. This may have an impact on the assortment strategy for the wine category. However, there could be certain group of customers for whom the wine category or products are a preference. These customers may be high value or regular customer who maybe few in numbers but prefer shopping in this category. In a scenario where the wine assortment or inventory is not well managed it could have a significant impact on the buying decision of these set of customers. This may cause the customer switching to competitor or abandoning the category purchase. Therefore in such scenario were a category of products who although may be have less sales figure but have a significant impact on consumer behavior. This leads us to a question that is there a way to know such categories and use this category level information to efficiently manage assortment. These are some of the hypothesis or question that I aim to answer using my research and come up with a method by which we can identify products that help build customer relationship and use it as an input while making assortment planning. And to test that adopting such a approach makes sense, I have proposed a hybrid assortment and compared it with a general assortment strategy.

Its a well-known fact that for most business the cost of customer acquisition is higher than the cost of customer retention (Min et al.; 2016). That is, it easier for business to get more business out of existing customer than from new customers. Today retailers and business owners depend upon product uptake and sales margin data to decide product assortment and stocking. Category, purchase managers have no data-driven insight on the product preference of the existing customers and what changes need to be adapted to cater to the demands of the customers. Also, with multiple products, categories, and SKUs, retailers have no clue which of them has a relatively higher weight over the others.

Therefore, it has now become very important to understand the needs and requirements of the existing customers and provide them with the best available options not only in terms of price but also in product categories and variety. Also, is important to identify products not only by their features but also by their relative impact on a consumer behavior. Product categorization and then using it as input to manage assortment will help businesses to offer an improved assortment method which ultimately improves customer relationship.

The entire thesis could be broken into several steps. Each step has a definitive aim towards the completion of the final project. The sections are as follows: -

1. Unsupervised segmentation on the customer base to identify loyal or ideal sets of customers who have good customer relationship value.
2. Creating product categories and combining it with customer segments to understand a consumers product buying behavior.
3. Assigning minimum thresholds or weights to product categories based on step 2 and then using it as an input to manage assortment.

The purpose of undertaking this research study was to understand the product-customer relationship and then create groups or categories of the products. Each category or group would then have certain weights or minimum support threshold according to their relative impact. This information would then be used as an input variable for creating a hybrid assortment strategy. Therefore, with this study, we would have a framework or a hybrid method which integrates customer behavior along with product attributes into an assortment strategy which helps a retailer achieve pre-defined customer relationship management (CRM) goals. Therefore, this entire study could be classified as an integrated research study combining key elements of CRM, product analytics and assortment management.

2 Literature and Related Work

To implement this research a thorough examination and review of related research reports were done. With the advancement in data analytics and machine learning methods, there have been several types of research done in this domain which have acted as a guiding step towards my research.

The most initial study done to understand product customer relationship was done by Lariviere and Van den Poel (2004). The study is considered as a stepping stone for working on research which involves exploring and building the customer-product relationship. Several researchers have used this study to form the basis of their research project. The study involved understanding if there are certain product or categories which help reduce customer churn and how cross-selling such product or services could help reduce customer churn. The outcome of the study was there was a significant impact of product or services on customer behavior and how important it was to offer different product and services to the customer to influence customer behavior thereby reducing customer churn. (Wang et al.; 2018) from their study demostrated how product features have a significant impact on customer satisfaction level. The study involved building a logistic regression model to assess the impact of different product features on customer satisfaction level.

Hong et al. (2016) demonstrated how a customer buying behavior towards certain product categories could be impacted in an assortment setup involving sharing of common assortment and display space. The researchers conclude that in a scenario where shoppers are exposed to items of categories who are not necessarily correlated in consumption then there could be a negative impact on customer buying behavior. ter Braak et al. (2014) Created a mechanism for retail assortment planning by creating an optimal assortment planning for private label (PL) brands. The study was motivated by the recent trends in emergency of number of PL brands due to their low cost. The researchers proposed model was developed by performing a computer assisted survey on consumers. The study did not use existing available data of the retailers or results from pilot launch of PL brands. One of the key product attributes which greatly impacts a customer buying decision is the price of product. Choi et al. (2018) research demonstrates difference in consumer reaction to difference in pricing assortments. The researchers conducted a study and created high and low-level groups of customers based on their choice satisfaction. The behavior of both set of customers was studied when subjected to availability of an assortment at parity on non-parity prices. The researchers measured choice confidence of the customers to measure the impact of the study on their behavior. Therefore, these studies highlight the need of having a better product, category level understanding along with consumer buying behavior while managing assortment decisions.

Melnic (2016) demonstrated how customer loyalty is the success in of a retailer in retaining and building a long terms engagement with the customer. The researcher worked out different customer segments based on behavioral and demographic data . Hwang et al. (2004) worked on a customer segmentation research for a wireless communication company. The study done by them created set of customers based upon customer lifetime value. A current customer value and potential customer value was used as a basis on segmenting customers for an insurance company. (Verhoef and Donkers; 2001).

Hebblethwaite et al. (2017) Carried a study in order to understand change in customer behavior towards discontinuation of unavailability of certain products or SKUs. The study highlighted how customer could switch stores or could be forced to buy deferment. The study was performed on 3 different scenarios of product discontinuation or replacement. The impact of all these 3 scenarios was evaluated as part of the study. The key outcome of the study that there was impact on customer behavior as per different scenarios. Clearly the research highlighted the key role of certain products and SKUs on customer buying behavior. The researchers established that retailers need to have a data driven customer-oriented approach while replacing or discontinuing certain products to avoid customer dis-engagement. The study also helps understand this type of customer-product behavior could also be used to manage assortment and identify those key products which have a greater impact on customer buying behavior. The studies guide us how we can understand or identify customer segments using some of the most popularly used customer segmentation strategies as stated below:

1. Using customer value for segmentation. (Zeithaml et al.; 2001)
2. customer segmentation by considering customer value and other factors (e.g., customer value, uncertainty, churn probability, etc.) (Benoit and den Poel; 2009)
3. As per Hwang et al. (2004) of segmenting customers by using only customer value

components e.g., current value, potential value, loyalty, etc.

These literature studies highlight the impact of customer-product relationship and view products as an important factor on consumer behavior. Reviewing these literature gives a understanding on the key areas to focus while implementing the research. Some of these research studies also provide a good understanding on the technical methodology that could be implemented for this research.

3 Methodology

This research is modelled on the Cross Industry Standard Process for Data Mining' (CRISP-DM)methodology. CRISP-DM is a hierarchical approach for implementing a data mining project (Wirth and Hipp; 2000). The key steps involved in CRISP-DM methodology and how it relates to this research is stated as follows:

Business Understanding: - This is the initial most stage of the data mining project as per the CRISP-DM methodology. For this research topic the business understanding would deal with understanding certain key CRM parameters which are critical for the business. For different business the CRM parameters may be different and the method or time frame of measuring them would also be different. Therefore, at this stage we understand the key business goal and then proceed to implement the data mining techniques to achieve it.

Data understanding: - The next step of CRISP-DM methodology is data the understanding phase. For this thesis the data understanding step involved exploring the data to understand the attributes, size, dimensions etc. The data used of this thesis project is a transactional data with customer and product level attributes. A thorough understanding of the data was done to proceed with the next stage of the process.

Data Preparation: - The data preparation steps involve creating the customer and product master data for performing customer segmentation and product clustering. For both the activities a unique master table needs to be derived with corresponding transactional attributes. These attributes would then be used to perform segmentation and clustering. For association rule mining data would need to be created in 'transaction' format to mine for frequently appearing itemsets.

Data modelling:- With the creation of the data masters table, we proceed with the data modelling phase. The data modelling phase for this project involves performing customer segmentation, performing product categorization using clustering methods and then performing association rules mining by using methods like Apriori and Weighted association rule mining.

Evaluation: - The outcome of the modelling phase would be evaluated at this stage. This would involve evaluation of each stage of the project as well as the evaluation of the entire project outcome. Since the aim of the project is to create categories of products based on their impact on customer behavior and then using it to manage assortment. Therefore, the evaluation would involve evaluating if this approach results in a different assortment strategy over a general strategy.

Deployment: - As this project more focused on the research side. Therefore, deployment stage would not be part of this project. Instead, future scope and steps ahead for taking this research forward could be discussed.

The entire process or methodology for the project could be demonstrated using a project flow diagram (Rivo et al.; 2012).

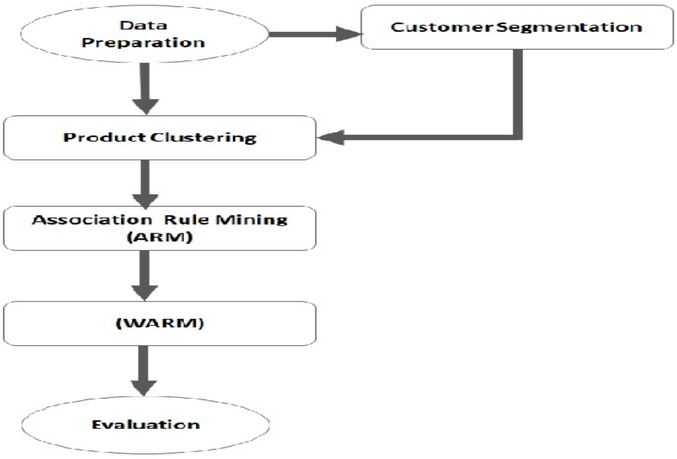


Figure 1: Project Flow

4 Implementation

The data for the project used is the online retail dataset available at UCI machine learning repository Chen et al. (2012). The dataset contains transactional data of more than 540,000 records with about 26,000 unique transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based online retail. Although this data is of an online retail store, for this research study the data is treated for a retail scenario. Data exploration and data cleaning was performed in data (Van den Broeck et al.; 2005). Few variables like date, stockcode etc. were formatted as per requirement. The entire project was executed using R programming and certain data exploration was done using excel.

The first step of the project deals with creating a customer and product master data in R. Master data represents most of the important entities of a companys business unit (Smith and McKeen; 2008). Some of the common master table include customer, product, store, location etc. The main characteristic of master data is that it is used by the entire company. Due to the organizational wide application and importance it important to define the master data unambiguously and maintain diligently across the organization (Ofner et al.; 2013). The master tables would be a unique datasets at a customer and product level. Several variables would be aggregated in the master tables which would be then used to analysis. Post creation of the master dataset we proceed with customer segmentation using clustering method. The method for performing the customer segmentation used in k-means clustering. K-means clustering is one of the most common unsupervised learning algorithm which tries to classify a given set of data points into certain number of cluster selected using k (Kanungo et al.; 2002). The algorithm tries to minimize the squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{i=1}^{c} i(||x_i - y_i||)^2 \tag{1}$$

Where,

$'\|x_i - y_i||'$ Denotes the Euclidean distance between y$_i andx_i$

$'c_i'$ is the total data points in the $i^{th} cluster$

$'c'$ is the number of cluster centers.

The most important step in k-means clustering algorithm is deciding the value of k. For this project the approach for selecting the value of k is done by elbow method. The elbow method is a iterative method which runs for different values of K. For each value of K the Sum of Square Error (SSE) is calculated. The SSE is then plotted in Y-Axis along with K number of clusters in the X-Axis. The optimal value of k is selected at placed where there is a elbow type curve in the graph plot i.e. the value of k post which the variation in SSE becomes constant.

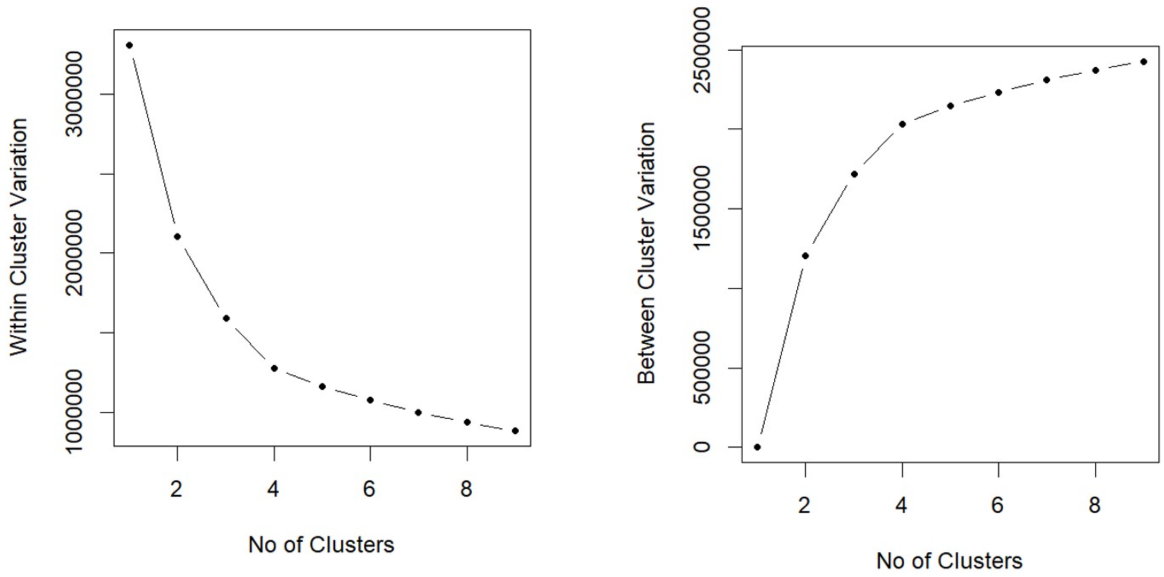


Figure 2: No of Clusters by Elbow Method

As seen the above fig the optimal value of k is selected as 4 using the elbow curve interpretation of the graph (Bholowalia and Kumar; 2014). The k-means algorithm was executed on the database. To graphically visualize clusters, we need to plot clusters against each variable. By plotting the clusters, we can see whether there is a need to merge or break clusters depending if there is an overlap of clusters or not. But since so many dimensions are hard to plot we would create principal components and plot the clusters against the first two PCs. As seen from the above 2-dimensional and 3-dimensional graphs the clusters look good and can be mapped back to the original data set to analyze the clusters using the clustering variables.

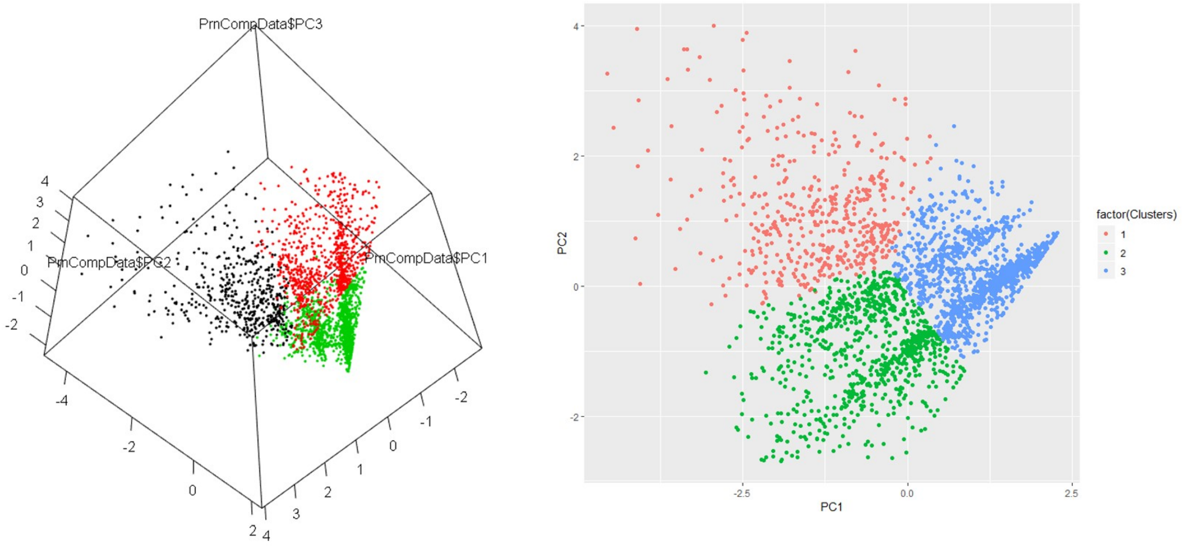
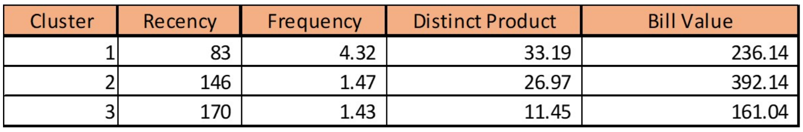


Figure 3: Cluster distribution

Post assigning the customer to respective clusters. The variable wise distribution along the clusters is calculated. This helps us to get a better understanding of the segments using the variables used for clustering.



Cluster	Recency	Frequency	Distinct Product	Bill Value
1	83	4.32	33.19	236.14
2	146	1.47	26.97	392.14
3	170	1.43	11.45	161.04

Figure 4: Variable Wise Cluster Analysis

As seen from the above table Cluster 1 is the most important cluster followed by cluster 2. Cluster 1 normally consists of Star customers with the best mean values across all variables. Cluster 2 consists of Loyal customer with an average mean value across the variables. Cluster 3 could be defined as the problem cluster.The customers in this clusters are mostly inactive (high recency) and with very low average frequency. For any retailer its important to move customers out of this cluster into better clusters. Several strategies like promotions, campaigns, offer etc. could be applied to improve the transactional behavior of these customers. However, like any other business its very important for the business to avoid customers from good clusters shifting or downgrading into poorer clusters. This indirectly would indicate shift in customer loyalty. Therefore, this project deals with increasing customers movement to better segments and avoiding downgrading of customer among segments by using a product focused approach. To implement this, we now need to understand the product buying behavior of these customers and try to establish a customer product relationship using the segment information. Therefore, we now proceed to understand difference in product buying behavior across the different customer segments. Since clusters 1 2 are more engaged customers in terms of the recency, frequency, moentary (RFM) metrics they could have some different product buying behavior as compared to cluster 3. This product level information becomes quite critical as it not only helps maintain engagement with good customers but can also be used to improve engagement from customers of other segments.

The next stage of the project deals with working on the product customer relationship and trying to understand the buying behavior of customer belonging to good clusters. For product clustering 2 clustering methods were tested and the best of the two was used for the final implementation. The techniques used were k-means clustering and fuzzy c-means clustering method. Fuzzy c-means (FCM) is another common type of unsupervised learning method to divide a data point into different segments or cluster. The difference in c-means clustering over other clustering method is that it allows a data point to belong to one or more cluster. (Dunn; 1973),(Peizhuang; 1983). FCM algorithm is also termed as soft clustering as a data point belonging to a certain cluster would have higher degree towards it centriod as compared to a centroid of another cluster. Like k-means FCM also tries to minimize the objective function subject to membership values of the data point to that cluster.

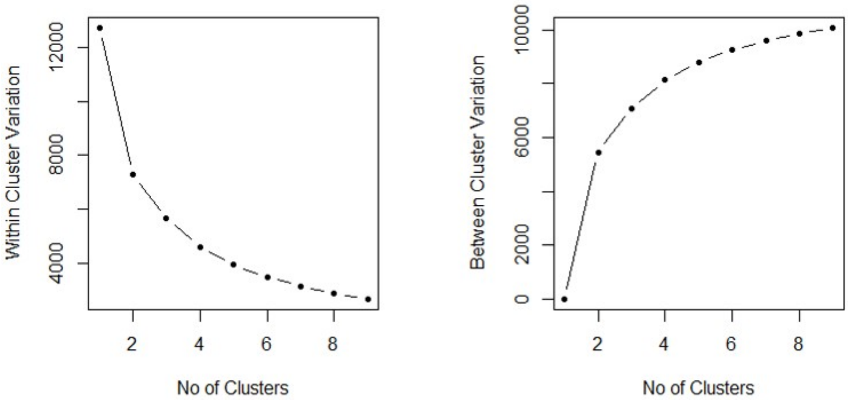


Figure 5: No of Cluster by Elbow method

By observing the above graph and using the eblow (Bholowalia and Kumar; 2014) method the no of cluster was selected as 6 and clustering algorithms were executed. With values k=6 and C=6

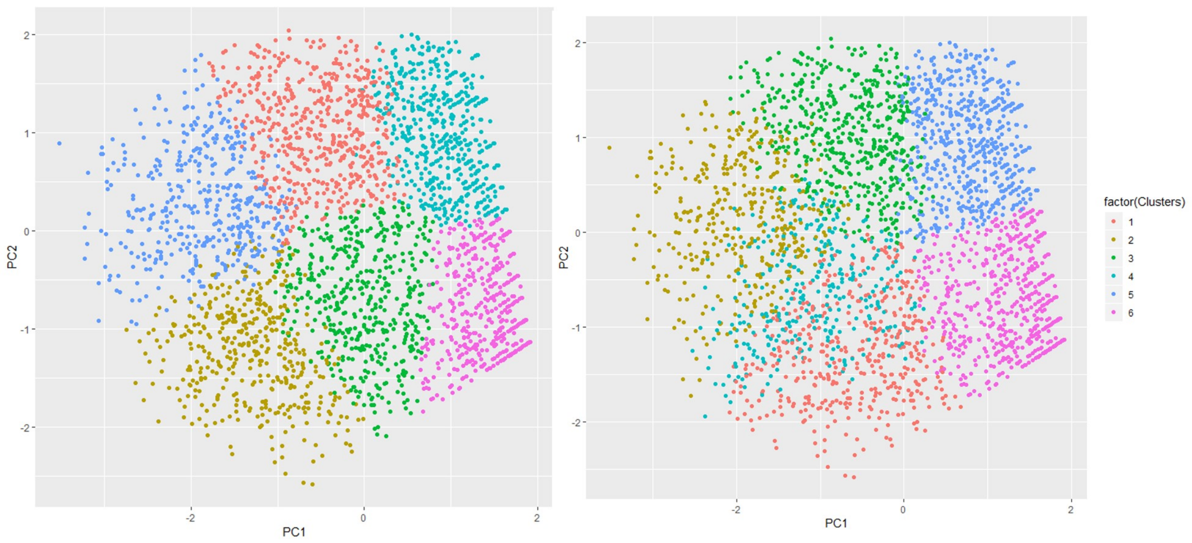
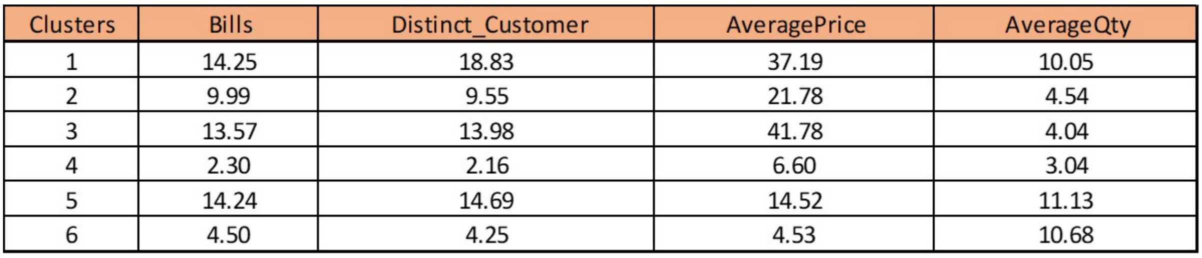


Figure 6: Fuzzy C-means vs K-means

As seen from the above visualization, Fuzzy C-means method has a better cluster split as compared to k-means clustering. Therefore, Fuzzy C-means method was selected as the final clustering method and the products were clustered.



Clusters	Bills	Distinct_Customer	AveragePrice	AverageQty
1	14.25	18.83	37.19	10.05
2	9.99	9.55	21.78	4.54
3	13.57	13.98	41.78	4.04
4	2.30	2.16	6.60	3.04
5	14.24	14.69	14.52	11.13
6	4.50	4.25	4.53	10.68

Figure 7: Cluster Analysis

As seen from the above clusters analysis. Cluster 1, 3 and 5 are the best product clusters in terms of mean average values of the variables. More customers have tired or brought these products. Also, the average prices of these products are also higher.

Post completion of customer and clustering the customer base was segmented into 3 clusters and the entire product range into 6 clusters. Now these cluster information was populated back into the transaction database to be used to find frequently occurring combinations in the transactions. Since cluster 1 (Star Customer) was the best customer segment in terms of RFM parameters it would be interesting to see the product buying behavior of these customers. To check the frequently brought category as per customer segment Apriroi algorithm would be used on the transactions data. Apriori algorithm is one of the most commonly used frequent set mining. The algorithm was proposed by Agrawal et al. (1994) to work on transactional databases to mine frequently occurring item combinations. For this study the aim would be to mine frequently customer segment with product categories.

Association rule mining (ARM) is used to create a set of rules which is used to denote relationship among the items. One of the most commmon way of denoting assocation rule is as {bread, milk} → {egg} which translates as if bread and milk is brought together than egg is also most likey to be brought as well. Such type association rule is termed as a market basket rule. Association rules could also be created for item pairs as {bread} →{milk}. However the rules in association rule mining of Apriori algorithms are dependent upon support and confidence. Support of an itemset denotes how frequent is the item in the data. Support for an itemset is given by the formula:

$$\text{Support(X)}=\frac{Count(X)}{N}$$

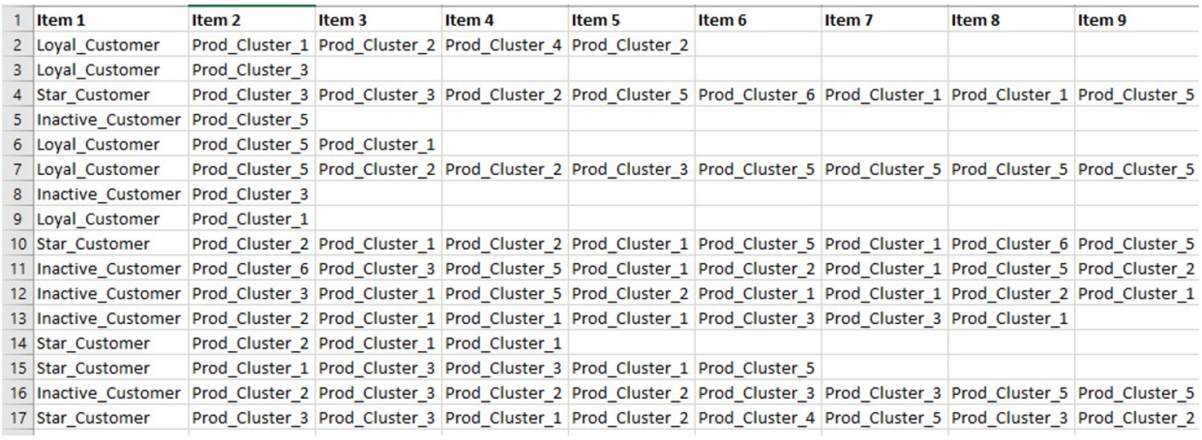
Where, Ń denotes the total number of transactions in the database. and (X) denotes the total transactions where itemset X appears.

Similarly, Confidence is defined as a predicitive power for a rule. It is given by the formula:

$$\text{(X}\rightarrow\text{Y)}=\frac{Support(X,Y)}{Support(X)}$$

Therefore, confidence is an indicator of the proportion of transactions where the of item X may result in the presence of item Y. Generally rules with high support and high confidence are termed as strong rules. Mostly cut-offs are taken to examine certain rules to study and understand possible combination of items.

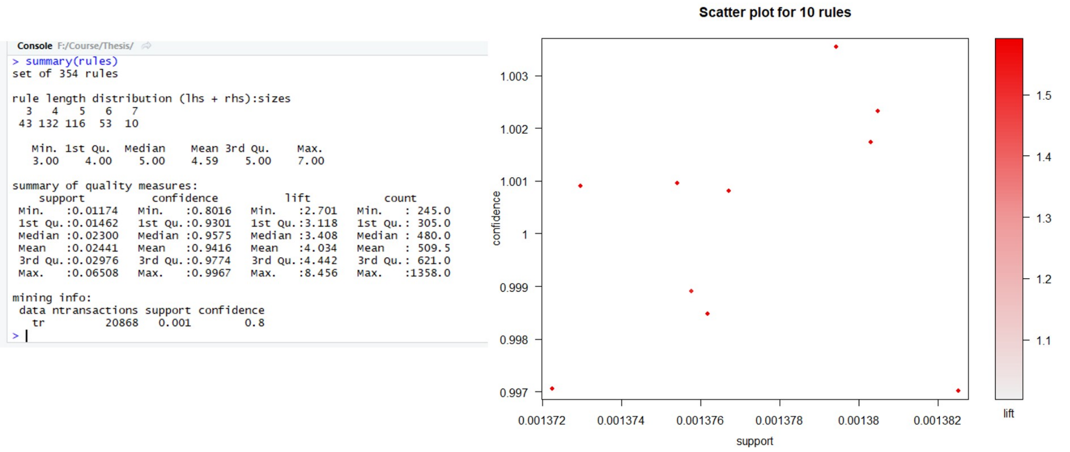
In this thesis the combination between customer segment and product cluster is tested. The data is created in a transactional data format. Customer clusters and product clusters now form part of the transactional data.



1	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
2	Loyal_Customer	Prod_Cluster_1	Prod_Cluster_2	Prod_Cluster_4	Prod_Cluster_2				
3	Loyal_Customer	Prod_Cluster_3							
4	Star_Customer	Prod_Cluster_3	Prod_Cluster_3	Prod_Cluster_2	Prod_Cluster_5	Prod_Cluster_6	Prod_Cluster_1	Prod_Cluster_1	Prod_Cluster_5
5	Inactive_Customer	Prod_Cluster_5							
6	Loyal_Customer	Prod_Cluster_5	Prod_Cluster_1						
7	Loyal_Customer	Prod_Cluster_5	Prod_Cluster_2	Prod_Cluster_2	Prod_Cluster_3	Prod_Cluster_5	Prod_Cluster_5	Prod_Cluster_5	Prod_Cluster_5
8	Inactive_Customer	Prod_Cluster_3							
9	Loyal_Customer	Prod_Cluster_1							
10	Star_Customer	Prod_Cluster_2	Prod_Cluster_1	Prod_Cluster_2	Prod_Cluster_1	Prod_Cluster_5	Prod_Cluster_1	Prod_Cluster_6	Prod_Cluster_5
11	Inactive_Customer	Prod_Cluster_6	Prod_Cluster_3	Prod_Cluster_5	Prod_Cluster_1	Prod_Cluster_2	Prod_Cluster_1	Prod_Cluster_5	Prod_Cluster_2
12	Inactive_Customer	Prod_Cluster_3	Prod_Cluster_1	Prod_Cluster_5	Prod_Cluster_2	Prod_Cluster_1	Prod_Cluster_1	Prod_Cluster_2	Prod_Cluster_1
13	Inactive_Customer	Prod_Cluster_2	Prod_Cluster_1	Prod_Cluster_1	Prod_Cluster_1	Prod_Cluster_3	Prod_Cluster_3	Prod_Cluster_1	
14	Star_Customer	Prod_Cluster_2	Prod_Cluster_1	Prod_Cluster_1					
15	Star_Customer	Prod_Cluster_1	Prod_Cluster_3	Prod_Cluster_3	Prod_Cluster_1	Prod_Cluster_5			
16	Inactive_Customer	Prod_Cluster_2	Prod_Cluster_3	Prod_Cluster_2	Prod_Cluster_2	Prod_Cluster_3	Prod_Cluster_3	Prod_Cluster_5	Prod_Cluster_5
17	Star_Customer	Prod_Cluster_3	Prod_Cluster_3	Prod_Cluster_1	Prod_Cluster_2	Prod_Cluster_4	Prod_Cluster_5	Prod_Cluster_3	Prod_Cluster_2

Figure 8: Transactional dataset for ARM

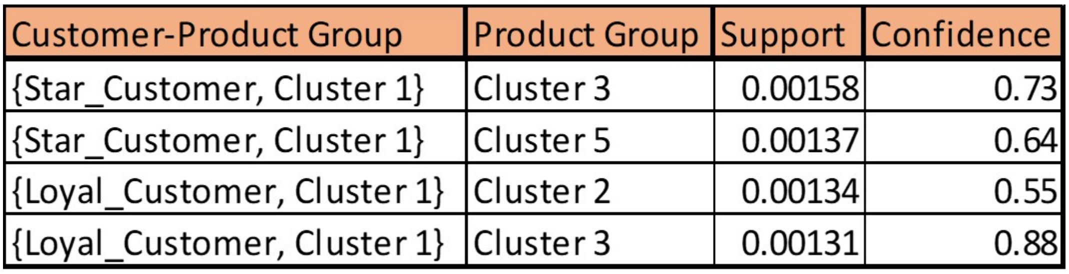
The next step involves creating the association rules and then plotting the rules graphically to visualize them. With changing the support and confidence levels different rules can be created for the database.



```
Console F:/Course/Thesis/
> summary(rules)
set of 354 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6   7 
 43 132 116  53  10 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   3.00    4.00    5.00    4.59    5.00    7.00 

summary of quality measures:
    support           confidence          lift            count       
 Min.   :0.01174   Min.   :0.8016   Min.   :2.701   Min.   : 245.0  
 1st Qu.:0.01462   1st Qu.:0.9301   1st Qu.:3.118   1st Qu.: 305.0  
 Median :0.02300   Median :0.9575   Median :3.408   Median : 480.0  
 Mean   :0.02441   Mean   :0.9416   Mean   :4.034   Mean   : 509.5  
 3rd Qu.:0.02976   3rd Qu.:0.9774   3rd Qu.:4.442   3rd Qu.: 621.0  
 Max.   :0.06508   Max.   :0.9967   Max.   :8.456   Max.   :1358.0  

mining info:
 data ntransactions support confidence
   tr         20868   0.001        0.8
> |
```

Figure 9: Rules Summary and Graph

The association rules as per customer and product segment was sorted as per confidence.

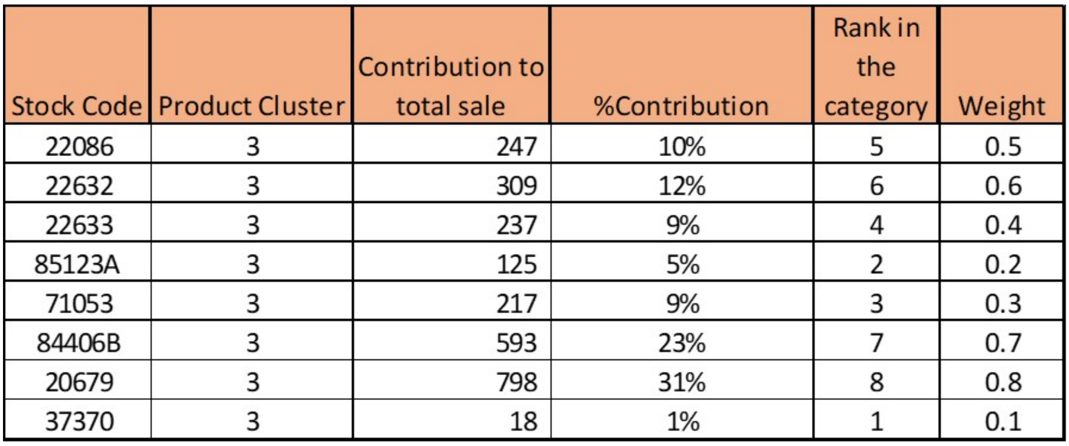


Customer-Product Group	Product Group	Support	Confidence
{Star_Customer, Cluster 1}	Cluster 3	0.00158	0.73
{Star_Customer, Cluster 1}	Cluster 5	0.00137	0.64
{Loyal_Customer, Cluster 1}	Cluster 2	0.00134	0.55
{Loyal_Customer, Cluster 1}	Cluster 3	0.00131	0.88

Figure 10: Customer-Product Rules

From the above figure its clear that Star customers who purchase prodcuts from cluster 1 are more likely to purchase products from cluster 3 and also from cluster 5. Similarly, loyal customer who purchase product from cluster 1 are more likely to purchase product from cluster 2 and cluster 3. Since cluster 1 is the best product cluster, most customer segment purchase those products and hence these products have a higher support among other clusters. From this outcome its evident that all though product cluster 1 has best sales metrics and is popular product category among customers, there are other categories of product which are also preferred by customers. It is also very clear from the product clustering that certain products due to their high sales metrics are part of this top cluster. Similarly, as per association rule mining results, this product category items will have higher frequency among top rules than other items of different categories. Therefore during any assortment or association rule mining activities these category of products would have high support values. Most ARM strategies involve exploring only the top rules as per the confidence due to this certain products may suffer from lower support levels and may not be part of the top rules. Therefore, in order to overcome this, minimum threshold levels is assigned to other category of products.

After assigning minimum threshold values or weights to product clusters, association rule mining (ARM) is again executed on the transactional database. The type of ARM used in this step is termed as Weighted Association Rule Mining or WARM. Since minimum threshold is assigned to certain products categories therefore this process can be considered as weight assignment method. **?** worked on improving weighted association rules technique for mining frequent itemset in a transactional database. The important characteristic of WARM it tries to maintain a balance between the weights of the items and the support of the itemset. In this study the method for assigning weights was done by using products contribution to the category. An example for this method could be seen in the below table.



Stock Code	Product Cluster	Contribution to total sale	%Contribution	Rank in the category	Weight
22086	3	247	10%	5	0.5
22632	3	309	12%	6	0.6
22633	3	237	9%	4	0.4
85123A	3	125	5%	2	0.2
71053	3	217	9%	3	0.3
84406B	3	593	23%	7	0.7
20679	3	798	31%	8	0.8
37370	3	18	1%	1	0.1

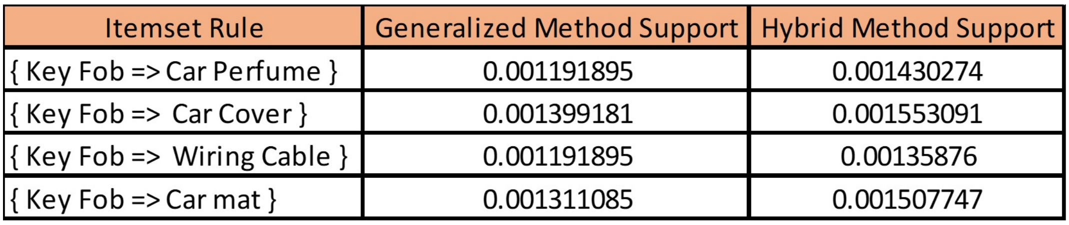
Figure 11: Sample weight assignment for cluster 3 products

Post assigning weights to the products, WARM was executed on the transactional database along with a normal apriroi algorithm on the transactional database.

5 Evaluation

Proper Evaluation of the results and outcomes is a key to an research outcome. During the implementation stage a number of times the outcomes were evaluated to understand the outcome of the activity. Successfully evaluation leads to successful implementation which paves way towards the next stage of the project.

The aim of this entire thesis is to evaluate or test the hypothesis that certain products due to their impact on customer behavior need to be treated differently. Post product clustering, mining frequent customer-product rules and performing WARM the project is finally evaluated to test if the Hybrid strategy yields different results over a generalized strategy.



Itemset Rule	Generalized Method Support	Hybrid Method Support
{ Key Fob => Car Perfume }	0.001191895	0.001430274
{ Key Fob => Car Cover }	0.001399181	0.001553091
{ Key Fob => Wiring Cable }	0.001191895	0.00135876
{ Key Fob => Car mat }	0.001311085	0.001507747

Figure 12: Generalized Vs Hybrid method

The above table represents the change in support for a item KEY FOB using the hybrid method in this research project. The top 5 rules for the item KEY FOB was evaluated using both the general and hybrid approach as proposed in this thesis. There could be significant impact on the association rules due to change in the support level of certain items sets. In large transactional datasets there are hundreds of different items which may have high confidence but due to lower support value may not be part of top rules. For e.g. {KEY FOB} →{CAR PERFUME} may have a very low support value but a high confidence level. i.e. customers who buy KEY FOB are very likely to buy CAR PERFUME and such customers maybe premium star customers. Due to low support value such rules may not be part of top rules.

6 Conclusion Future Work

The main aim of this study was to test the hypothesis that different products have different impact on customer behavior. Therefore, these products should be categorized as per their impact and ultimately be used as input in a hybrid assortment strategy. To test this hypothesis, customer-product relationship was established using unsupervised clustering along with association rules mining. The output of this process was assigned weights and used as input in a weighted association rule mining (WARM) method. The result observed for an item showed difference in support level for the item over a generalized assortment method. Therefore by using this hybrid strategy new rules which earlier where not significant could be mined. By creating product clusters or categories and assigning weights to impact full products retailers can easily identify such products and can then tailor their assortment strategy using this hybrid approach.

This project provides an abundance of opportunity to explore customer-product relationship further and categorize of create set of products which may have an impact on customer behavior. Product categorization is a field which is going to throw up a number of challenges as time passes by. With new product development, changing customer demographics, a products attributes and characteristics would keep on changing. Retailers will always struggle to know which are the key products which directly or indirectly influence customer behavior. Therefore, future researchers can deep dive more into understanding customer-product relationships to create set or categories of products. Deep learning methods would be a great way in order to achieve this. The new categorise developed could be tested using a variety of hybrid assortment techniques or test launches by the retailers. The ultimate aim of this study would be create a grading method or framework which a retailer can use to grade his products based on its overall attribute.

References

Agrawal, R., Srikant, R. et al. (1994). Fast algorithms for mining association rules, *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215, pp. 487–499.

Benoit, D. F. and den Poel, D. V. (2009). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services, *Expert Systems with Applications* **36**(7): 10475 – 10484.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417409000712*

Bholowalia, P. and Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn, *International Journal of Computer Applications* **105**(9).

Chen, D., Sain, S. L. and 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): 197–208.
URL: *https://doi.org/10.1057/dbm.2012.17*

Choi, C., Mattila, A. S. and Upneja, A. (2018). The effect of assortment pricing on choice and satisfaction: The moderating role of consumer characteristics, *Cornell Hospitality Quarterly* **59**(1): 6–14.
URL: *https://doi.org/10.1177/1938965517730315*

Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters, *Journal of Cybernetics* **3**(3): 32–57.
URL: *https://doi.org/10.1080/01969727308546046*

Hebblethwaite, D., Parsons, A. G. and Spence, M. T. (2017). How brand loyal shoppers respond to three different brand discontinuation scenarios, *European Journal of Marketing* **51**(11/12): 1918–1937.
URL: *https://doi.org/10.1108/EJM-08-2016-0443*

Hong, S., Misra, K. and Vilcassim, N. J. (2016). The perils of category management: The effect of product assortment on multicategory purchase incidence, *Journal of Marketing* **80**(5): 34–52.
URL: *https://doi.org/10.1509/jm.15.0060*

Hwang, H., Jung, T. and Suh, E. (2004). An ltv model and customer segmentation based on customer value: a case study on the wireless telecommunication industry, *Expert Systems with Applications* **26**(2): 181 – 188.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417403001337*

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. and Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(7): 881–892.

Lariviere, B. and Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services, **27**.

Melnic, E. L. (2016). How to strengthen customer loyalty, using customer segmentation?, *Bulletin of the Transilvania University of Brasov. Economic Sciences. Series V* **9**(2): 51.

Min, S., Zhang, X., Kim, N. and Srivastava, R. K. (2016). Customer acquisition and retention spending: An analytical model and empirical investigation in wireless telecommunications markets, *Journal of Marketing Research* **53**(5): 728–744.
URL: *https://doi.org/10.1509/jmr.14.0170*

Ofner, M. H., Straub, K., Otto, B. and Oesterle, H. (2013). Management of the master data lifecycle: a framework for analysis, *Journal of Enterprise Information Management* **26**(4): 472–491.
URL: *https://doi.org/10.1108/JEIM-05-2013-0026*

Peizhuang, W. (1983). Pattern recognition with fuzzy objective function algorithms (james c. bezdek), *SIAM Review* **25**(3): 442.

Rivo, E., de la Fuente, J., Rivo, Á., García-Fontán, E., Cañizares, M.-Á. and Gil, P. (2012). Cross-industry standard process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management, *Clinical and Translational Oncology* **14**(1): 73–79.
URL: *https://doi.org/10.1007/s12094-012-0764-8*

Smith, H. A. and McKeen, J. D. (2008). Developments in practice xxx: master data management: salvation or snake oil?, *Communications of the Association for Information Systems* **23**(1): 4.

ter Braak, A., Geyskens, I. and Dekimpe, M. G. (2014). Taking private labels upmarket: Empirical generalizations on category drivers of premium private label introductions, *Journal of Retailing* **90**(2): 125 – 140. Empirical Generalizations in Retailing.
URL: *http://www.sciencedirect.com/science/article/pii/S0022435914000049*

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R. and Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities, *PLOS Medicine* **2**(10).
URL: *https://doi.org/10.1371/journal.pmed.0020267*

Verhoef, P. C. and Donkers, B. (2001). Predicting customer potential value an application in the insurance industry, *Decision Support Systems* **32**(2): 189 – 199. Decision Support Issues in Customer Relationship Management and Interactive Marketing for E-Commerce.
URL: *http://www.sciencedirect.com/science/article/pii/S0167923601001105*

Wang, Y., Lu, X. and Tan, Y. (2018). Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines, *Electronic Commerce Research and Applications* **29**: 1 – 11.
URL: *http://www.sciencedirect.com/science/article/pii/S1567422318300279*

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining, Citeseer.

Zeithaml, V. A., Rust, R. T. and Lemon, K. N. (2001). The customer pyramid: Creating and serving profitable customers, *California Management Review* **43**(4): 118–142.
URL: *https://doi.org/10.2307/41166104*