TO CITE THIS PAPER, PLEASE USE:

Nie, D., Cappellari, P. and Roantree, M. (2020), "A methodology for classification and validation of customer datasets", Journal of Business & Industrial Marketing, Vol. 36 No. 5, pp. 821-833. https://doi.org/10.1108/JBIM-02-2020-0077

A Methodology for Classification and Validation of Customer Datasets *

Dongyun Nie[1] , Paolo Cappellari[2] , and Mark Roantree[1]

[1]Insight Centre for Data Analytics, School of Computing,
Dublin City University, Dublin 9, Dublin, Ireland.
[2]College of Staten Island
City University of New York, New York, USA.

Abstract

Purpose - The purpose of this research is to develop a method to classify customers according to their value to an organization. This process is complicated by the disconnected nature of a customer record in an industry such as insurance. With large numbers of customers, it is of significant benefit to managers and company analysts to create a broad classification for all customers.

Design/Methodology/Approach - The initial step is to construct a full customer history and extract a feature set suited to Customer Lifetime Value calculations. This feature set must then be validated to determine its ability to classify customers in broad terms.

Findings - Our method successfully classifies customer datasets with an accuracy of 90%. We also discovered that by examining the average value for key variables in each customer segment, an algorithm can label the group of clusters with an accuracy of 99.3%.

Research limitations/implications - Working with a real-world dataset, it is always the case that some features are unavailable as they were never recorded. This can impair the algorithm's ability to make good classifications in all cases.

Originality/Value - We believe that this research makes a novel contribution as it automates the classification of customers but in addition, our approach provides a high level classification result (recall and precision identifies the best cluster configuration) and detailed insights into how each customer is classified by two validation metrics. This supports managers in terms of market spend on new and existing customers.

Keywords - Customer Lifetime Value, Customer Segmentation, Clustering, Unsupervised Learning.

Paper Type - Research paper

*This research work was funded by Science Foundation Ireland under grant numbers: SFI/12/RC/2289 and SFI/12/RC/2289-P2.

1. Introduction

CRM (Customer Relationship Management) is a well established strategy to support companies in their management of customer relationships across 4 dimensions: Customer Identification (Customer Segmentation); Customer Attraction (Direct Marketing); Customer Retention (Loyalty Program); and Customer Development (Customer Lifetime Value) (Ngai et al. (2009)). One of the major goals of CRM is to maximize the *Customer Lifetime Value* (CLV) for the purpose of supporting long term business investment (Ling and Yen (2001)). *Customer Lifetime Value* is a measure that focuses on predicting the net profit that can accrue from the future relationship with a customer (Di Benedetto and Kim (2016)). This metric can be calculated by recording the behaviors of the customer over the longer term and thus, help to build a customized business strategy. It has been a popular research topic, addressed by researchers in different ways, for example, Formulaic CLV (Berger and Nasr (1998); Blattberg and Deighton (1996); Reinartz and Kumar (2000)) and Probability Model CLV (Pfeifer (2005); Sohrabi and Khanlari (2007); Fader et al. (2004)).

Customer segmentation is regarded as a natural process to help managers and company analysts to classify customers and plan market investment strategies such as direct sales. As such, it has been widely adopted by industry planners (Hiziroglu and Sengul (2012); Khajvand et al. (2011); Teichert et al. (2008)). Moreover, it plays a critical role in the development of a company's position by combining product differentiation and marketing segmentation to provide resources, objectives and competences to the company.

Various customer segmentation methods have been developed (Swanson Sr (2011)), using clustering algorithms (Sánchez-Hernández et al. (2013)), different approaches to classification (Kim et al. (2006)), and regression methods (Hwang et al. (2004)), where each approach was selected to meet a specific goal (Müllensiefen et al. (2018); Lu et al. (2019); Ahani et al. (2019)). In general, some form of clustering has been the most popular method in customer requirements, queries, reports (Roantree and Liu (2014)) and segmentation projects (Wu and Chou (2011)).

1.1. Problem Description and Motivation

The research presented in this paper arises from a collaboration with a large insurance company based in Ireland. The problem at hand was to generate CLV metrics for a customer base from half a million policies across various insurance types. In line with many other organizations and for many reasons, their customers had fragmented records, many with multiple policies, where they appeared as distinct customers. This problem is exacerbated by the fact that there is no clear method to unify customer records. Furthermore, the main issue with most of the research into Customer Lifetime Value is that it is highly theoretical in nature and thus, lacking validation using real world data. Under normal conditions, the performance of clustering algorithms relies on the variable selection strategy. Thus, to deliver optimal segmentation of customers, the availability and selection of the correct variables is crucial. However, in many cases, variables that are available are not suited to machine learning algorithms with the identification and preparation Han et al. (2011) of the dataset seen as a complex process. Another critical task requires

the transformation of variables (Components) (King and Jackson (1999); Westad et al. (2003); Granato et al. (2018); Scriney et al. (2019); Yun et al. (2019)) using the appropriate normalization methodology (Milligan and Cooper (1988)). Our research is based on a problem that is common to many retailers: computing CLV values where customer records are separated across different point of sales with no clear method to combine them. In summary, a single holistic record does not exist for each customer and in many cases, there exists no identifiable attribute that can be used to combine a fragmented customer history. However, the formation of a complete customer history is crucial to extracting a dataset suitable for machine learning and predictive algorithms.

The ultimate goal of this research is to provide a CLV ranking for each customer in our collaborator's dataset and in previous research Nie and Roantree (2019), we performed the record linkage process to deliver unified customer records. However, many of the parameters required for CLV calculations are missing from this dataset and the final process will require a level of imputation to generate those missing values. This paper focuses on the step between record linkage and imputation, the goal of which is to construct a dataset ready for imputation by classifying customers into broad groups (good, bad, average) and validating these classifications.

1.2. Contribution and Paper Layout

As the company dataset is untrained, *clustering* is necessary to *group* customers and where necessary to remove outliers. However, once clustering has been completed, it is necessary to label or *classify* each customer group. Company managers and analysts have a finite budget spend on customer marketing and understanding the category to which a customer belongs, supports the decision on spending for that group of customers. This high level clustering of customers will also be crucial when validating the final CLV results in subsequent research.

The contributions which are novel to this research can be articulated as:

- A feature extraction process to generate the variables suited to data mining.
- A detailed approach to auto-classification of customer records that yields high results and informs managers in terms of allocating their marketing budget.
- Using different variable extraction methods and two sets of validation metrics, we show how two different sets of primary components were unable to outperform the full feature set, while highlighting where a reduced feature set may be optimal. This is important for managers and analysts in *understanding* both the process and final outcome in creating customer groupings.
- Customer groups undergo a robust validation using both subjective and objective evaluations methods.

Paper Structure. The remainder of this paper is structured as follows: in Section 2, we present a review of related research in this area; in Section 3, we describe the methodology for feature extraction (Guyon et al. (2008)) from the unified customer records, describe the feature set, and discuss different feature selection policies (Dash and Liu

(1997)); in Section 4, we describe our validation in terms of the experiments, results and findings; and finally, Section 5 contains our conclusions.

2. Related Research

Marketing researchers have been studying customer segmentation for decades. In Kara and Kaynak (1997), the authors attempted to place customers into different marketing groups where they used traditional segmentation methods such as normative segmentation, niche marketing, micro marketing, database marketing, relationship marketing, and mass customization. For the authors, normative segmentation meant clustering customers using the average *within-group* similarity (Wind (1978)). Their results demonstrated that finer segmentation (tailoring needs for the *individual* customer) was more successful than traditional segmentation. It increased customer retention and loyalty, with a higher competitive advantage for the company that had a flatter organizational structure. While this work developed a number of interesting hypotheses, there was no attempt to validate these ideas whereas in our research, we focus on a large number of experiments using our industrial collaborator's dataset to validate our findings.

A conceptual model for customer ranking, based on principal components, was presented in Parkhimenka et al. (2017). This approach comprised multiple steps: Variable Definition; Data Extraction; Variable Scaling; Principle Component Analysis; specification and development of ranking functions; and ultimately, grouping customers. Unlike our approach, however, their research was focused entirely on the *theoretical* aspect of customer ranking. While we developed a similar method, we also employed a real-world dataset and validated using a large number of experiments.

The researchers in Müllensiefen et al. (2018) present two case studies using survey data collected from subjects with different socio-demographic characteristics. Their primary goal was to identify the contribution of variables to the customer segmentation process. In their experiments, their clustering algorithms delivered between 2 and 10 clusters with an analysis as to how socio-demographic and personal variables performed in the different cluster configurations. Interestingly, their conclusions were that personality variables were more important for accurate market segmentation. While we also employ clustering algorithms to segment the customer dataset, we also provide a full Extract-Transform-Load framework where data is acquired from operational databases, integrated (customer record linkage), and transformed for machine learning.

The researchers in Sánchez-Hernández et al. (2013) presented a classification selection method by five fuzzy criteria with aggregated Ordered Weighted Averaging (OWA) and Multi-Criteria Decision Making (MCDM) system for customer segmentation. The OWA operator allows the implementation of the concept of *fuzzy majority* in the aggregation phase by means of a fuzzy linguistic quantifier (Zadeh (1983)). The five fuzzy criteria are: the useful number of classes; balanced classes; coherent classification; dependency and accuracy criterion of the predictive model. The aggregated OWA was decided on three steps: re-order the input arguments in increasing order; determine the weights for the operator in a proper way and then use OWA weights to aggregate the re-ordered arguments. They then applied identified classifications using an unsupervised method. They

evaluated a clustering method using three types of validation criteria: internal, external and relative. By using a marketing case study based on a B2B environment, their result showed that a cluster size of 3 delivers the most suitable marketing strategy. In terms of our research, our domain and dataset is different as our research requires a far longer time-frame to construct variables.

In Cibulková and Sulc (2018), the authors used hierarchical clustering methods (single-linkage, average-linkage and complete-linkage) to a travel dataset, in combination with different distance measures. They compared three hierarchical clustering methods in combination with five distance measures with the optimal k determined by low *within-cluster* variability (Rezankova et al. (2011)). The dataset used in this research is customer travel data with all variables either categorical or binary data and also, the dataset was far smaller to that used in our research. Moreover, the details of variable transformation and selection were not discussed.

Researchers such as Paschen et al. (2019); Martínez-López and Casillas (2013); Singh et al. (2019); Syam and Sharma (2018) put the focus on the need for including AI driven solutions in daily business operations, including decision making at the marketing and executive level. These papers provide a comprehensive, but generic (high-level), view of: the building blocks of AI systems, the interaction between these systems, and the advantages that market analysts and executives may have when some form of artificial intelligence is integrated in the organization's information systems. In comparison, our work goes more in-depth, and presents how some techniques from the AI realm can be applied to real-world data to obtain practical market insights to support managerial decision making.

Multistage market segmentation is addressed in a number of research projects, including Thomas (2016); Hillebrand and Biemans (2011); Kleinaltenkamp et al. (2012). These works investigate as to how managers can achieve a deeper competitive advantage when segmenting markets at different stages. In Hillebrand and Biemans (2011), the authors focus on the need to look beyond the direct customers in order to better understand downstream needs and thus, identify opportunities. In Kleinaltenkamp et al. (2012), researchers describe the challenges that a multistage market analysis may pose, while also discussing preconditions for it to succeed. The work in Thomas (2016) makes a further step in multistage market analysis by considering how to align the needs of the markets (and customers) at the different stages, so as to achieve a more cohesive, and advantageous, strategy to the market. The focus of our work, on the other hand, is in providing better customer management, which is different from trying to align multistage markets. Another major difference with our work is the depth and extent of the real-world scenario, as well as the fact that we are not making apriori assumptions in our classification, thus removing any potential bias.

In Fotiadis and Vassiliadis (2017), the authors perform a quantitative analysis of the relevance of CRM metrics in decision making. The study applies modern customer centered metrics, i.e. share and size of Wallet Malthouse and Blattberg (2005); Keiningham et al. (2005), recency, frequency, monetary value Fader et al. (2005), to a real-world dataset from the maritime shipping industry. The study suggests that each index can provide, directly or indirectly, critical performance measures to interested parties and the

management of each company within their industry. Similarly to Fotiadis and Vassiliadis (2017), we adopt RFM as an important metric in our work. However, in comparison to Fotiadis and Vassiliadis (2017) and the other cited works, the complexity of the real-world scenario and analysis we consider in this paper is of a greater scale, both in terms of the size of the data, and of the variables included in our study.

The Data Envelopment Analysis (DEA) method is the focus of many researchers, including Noorizadeh et al. (2013) and Shabanpour et al. (2017). These research efforts focus on evaluating the efficiency of decision making units (DMUs). The authors in Noorizadeh et al. (2013) root their approach in a variant of the BCC version of the cross-efficiency for the DEA method, where ratio values can be used in the evaluation of customers rankings. This approach supports customer evaluation in a multi criterial context, does not require decision makers to define weights for the process, and can accommodate both ratio and absolute numbers. In Shabanpour et al. (2017), the authors adopt an artificial neural network, in place of traditional regression methods, for forecasting the efficiency of future DMUs. Our work also focuses on managing customer value. However, while Noorizadeh et al. (2013) and Shabanpour et al. (2017) only deal with the scenario of verifying individual customer values, our research addresses customer segmentation, which is to organize customers in similarity groups. Moreover, the purpose of their case study is to show an example of the approach. Our work, instead, is applied to a real-world dataset with a much larger size.

Market segmentation and decision making is an active area of research that spans over a number of decades, and many other works exist that address this subject, including Powers and Sterling (2008); Sudharshan and Winter (1998); Merrilees et al. (1999), to mention a few more. Overall, when compared to the existing literature, the work proposed in this paper differentiates itself from others because of the unique combination of the data mining method adopted for customer segmentation, the validation method strategy, and the complexity of the real-world scenario and data used for the experiments.

3. Methodology

Insurance policies are purchased and renewed on a yearly basis, taking place over a variety of sales platforms, including phone, online website, store, and agent. A single person can sign for multiple policies, either as individual or as part of a family group. As a result, policies for the same customer can be stored in a different database depending on purchase type with the result that no unified customer records exist. The net result is that companies may not be aware that two different policies belong to the same customer. In previous work (Nie and Roantree (2019)), we addressed the linkage problem but this was merely step 1 in our overall approach. This section of the paper focuses on the extraction of a principal dataset and using these variables as input to the customer classification process.

3.1. Record Linkage

Data acquisition began with an export from company databases to create a dataset that was (insurance) *policy* centric. This was not useful for identifying a dataset of *good* customers

and thus, the goal of the Record Linkage step was to have a *customer* centric dataset. Record linkage is the task where algorithms are used to identify the same individual where client history is spread across multiple records often within the *same* dataset. In cases where unique identifiers are found, linking those records is a trivial task. However, there are very high numbers of individuals who cannot be matched as common identifiers do not exist across datasets. This step is not part of this research but was presented in Nie and Roantree (2019). In this earlier work, a customer matching approach was developed using a modified form of Agglomerative Hierarchical Clustering (AHC), together with a companion ruleset to detect and link components of the same customer record, and also clients with family members. The result was relatively fast matching while achieving high levels of accuracy. After this step, the dataset reduced from 500,859 records to 387,951 unified customer records, meaning 112,908 policies were matched to existing customers.

3.2. Data Transformation

While the recordset is now annotated with links, data instances are at the *policy* level. The purpose of the transformation process to create a dataset with one instance per customer and with variables that are appropriate for machine learning algorithms. Of the variables selected (and described below), a number are common to various research projects (Barnes et al. (2007); Lee and Philip (2010); Walsh et al. (2010); Christy et al. (2018); Müllensiefen et al. (2018)): Age, County (Address), Gender, and Margin. Given our access to raw data over a 6-year period, we used the opportunity to include and test a number of less commonly used variables: AdjAmt, AdjCT, CCNbr, CFNbr, Disct, Gap, Maxh, and YNbr (staying duration).

TABLE I
CLV Variable Descriptions

Variable	Description
AdjAmt	Cost associated with adjustments required by the customer.
AdjCT	Number of adjustments made by the customer.
Age	Customer Age.
CCNbr	Number of policies a customer holds.
CFNbr	Number of policies held by a customer's family member.
County	County code part of the customer's address.
Disc	Discount rate the customer receives, based upon each policy.
Gap	Indicates if the customer is a returning customer.
Gender	Customer Gender.
Margin	Net income generated by the customer.
Maxh	Maximum number of policies a customer held for any calendar year.
YNbr	Total number of years a customer held at least one policy.

In table I, we present the set of variables generated from the data transformation process. While some of the variables require no further description, there are some semantics associated with others, that benefit from closer inspection.

- **AdjAmt:** This variable represents the cost (in monetary terms) associated with adjustments a customer made during their tenure (AdjCT), e.g. a policy modification due to a change of car. This measure is highly correlated to customer retention in the insurance industry (Smith et al. (2000)): policy modifications are more likely to affect the premium (policy price). Where customers leave after a policy modification, AdjCT and the AdjAmt are very important.

- **CCNbr:** This variable is cumulative and highly correlated with *Margin*.

- **CFNbr:** The number of policies held by a customers family member. It is assumed that this measure has some impact on the customer's spending or retention behaviour.

- **Disct:** The discount rate the customer receives, based upon each policy. The discount is a rate calculated by $1-Value_p/Value_b$, where: $Value_b$ is a basic calculation on a policy's worth; and $Value_p$ represents the *actual* price they were charged. For each policy and year, a discount rate is applied. The Discount present in the dataset is an *average* rate based on whether a policy is held for multiple years or a client had multiple policies. This is a crucial variable in determining the retention and acquisition figures for a customer.

- **Margin:** The *net income* generated (purchased) by the customer. Margin is the total policy spend over the 6 years. This is a very common variable for segmentation and is the *Monetary* variable in RFM model.

- **Maxh:** The maximum number of policies a customer held for any calendar year. This is a measure to capture the customer's buying behaviour in the short term (a single year) and differs from *CCNbr*, which captures the entire customer record.

- **YNbr:** The total number of years a customer held at least one policy. The value measures customer loyalty for the company, which is a very important variable for CRM. This variable is the same as *Frequency* in the RFM validation model.

TABLE II
Descriptive Summary for Variables.

	AdjAmt	AdjCT	Age	CCNbr	CFNbr	County	Disct	Gap	Gender	Margin	Maxh	YNbr
mean	72.06	1.61	52.13	1.29	0.05	N/A	0.59	0.08	N/A	1336.41	1.16	2.57
std	300.08	2.34	16.47	0.65	0.23	N/A	0.01	0.39	N/A	1109.13	0.41	1.61
min	0	0	18	1	0	N/A	-4.44	0	N/A	33	1	1
25%	0	0	39	1	0	N/A	0.52	0	N/A	586	1	1
50%	0	1	51	1	0	N/A	0.63	0	N/A	996	1	2
75%	34	2	64	1	0	N/A	0.69	0	N/A	1756	1	3
max	33562	111	117	33	3	N/A	0.85	4	N/A	22696	22	6

Good practice when dealing with new datasets, is to perform a descriptive summary to obtain an overall picture of the data. Table II presents this summary for each of the 12 variables, where: row **mean** presents the average of that variable; **std** the standard deviation; **min** the minimum value; and **max** the maximum values for that variable. In

addition, **25%**, **50%** and **75%** represent the 25th, 50th, and 75th percentiles respectively. Finally, as the dataset reflects a period of 6 years, the maximum value of YNbr is 6. The variables County and Gender are categorical, as such they do not have aggregate statistics such as mean and standard deviation (std). This explains the presence of `N/A` (not applicable) values in table II.

From this analysis, we can see that all variables are numeric but with very different ranges of values. The difference (min to max) in Margin and AdjAmt is very high in comparison to all others. For some variables, the percentiles show little or no increase: AdjAmt, CCNbr, CFNbr, Gap, Maxh variables change only at the top 25th percentile (max). Given the broad range of variables, it is crucial that one variable does not overly influence algorithms and thus, the dataset was normalized.

3.3. Correlation Coefficient Analysis

The Correlation Coefficient as defined by Lawrence and Lin (1989) quantitatively describes the degree to which two variables are related. The correlation coefficient formula used in this work is presented in Equation 1, where: CC_{xy} represent the correlation value between the (pair of) variables x and y; s_x and s_y are the sample standard deviations; and s_{xy} is the sample co-variance.

$$CC_{xy} = \frac{s_{xy}}{s_x s_y} \tag{1}$$

Table III shows the correlation values between all variables with values ranging between 0 and 1. The higher the value for the correlation between two variables, the higher those two variables are related. For instance, the two most correlated variables are CCNbr and Maxh, because their correlation value, 0.79, is higher than all others; conversely, CCNbr has no relation with AdjAmt, since the correlation value is ≈0. Table III also has an additional column for All Correlation Coefficients (**ACC**): the values in this column compares a variable with *all* other variables. It provides a score for each variable by summing its correlation value to provide an overall ranking for variables.

TABLE III
Correlation Coefficient Results

	AdjAmt	AdjCT	Age	CCNbr	CFNbr	County	Disc	Gap	Gender	Margin	Maxh	YNbr	ACC
AdjAmt	1												0.99
AdjCT	**0.47**	1											2.04
Age	0.04	0.12	1										0.71
CCNbr	0.00	0.26	0.04	1									2.52
CFNbr	0.00	0.01	0.03	0.25	1								0.39
County	0.00	0.01	0.01	0.01	0.04	1							0.43
Disc	0.04	0.18	0.22	0.04	0.02	0.07	1						0.82
Gap	0.03	0.07	0.00	**0.33**	0.01	0.00	0.03	1					0.67
Gender	0.00	0.00	0.02	0.00	0.00	0.21	0.01	0.00	1				0.27
Margin	0.23	**0.39**	0.02	**0.45**	0.00	0.03	0.05	0.06	0.00	1			2.4
Maxh	0.07	0.19	0.08	**0.79**	0.02	0.01	0.09	0.05	0.00	**0.45**	1		1.98
YNbr	0.10	**0.33**	0.12	**0.34**	0.01	0.04	0.06	0.09	0.00	**0.71**	0.25	1	2.05

It can be said that variables exhibiting high correlation values carry the same information and are thus, redundant. Data from redundant variables do not improve the quality of the clustering results and should be removed. The correlation between two variables can

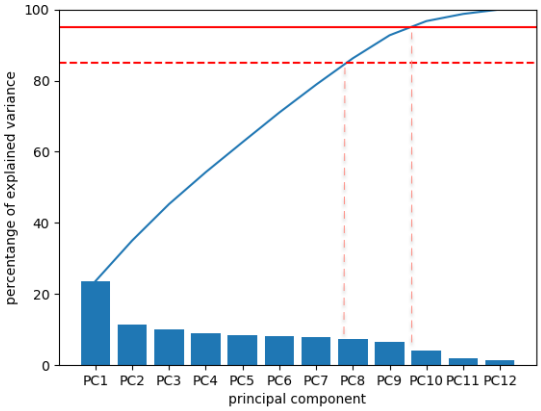


FIGURE 1: Variance of different Principal Components Explained

be defined as high when exceeding the 0.3 mark threshold, as per Asuero et al. (2006). Correlation values exceeding the set threshold are highlighted in table III, which in turn identify highly correlated variables in associated columns and rows. Using an algorithm to process the correlation values in table III, the following variables were removed: Margin, AdjCT, and CCNbr.

3.4. Principle Component Analysis

Principle Component Analysis (PCA) is predominantly used when we have a large set of variables. In brief, PCA will take the full set of variables and generate a set of *dimensions*, constructed from this original variable set. Thus, the goal of this process is to identify the best set of dimensions. As it is one of the more common approaches to variable selection, we use it as a baseline in our evaluations. Using PCA, we determined the importance of components by the *Variance Ratio* of each variable. The variance ratio illustrated in figure 1, shows the same components but in a format that emphasizes the *impact* of each PC on the overall percentage of the variance.

Specifically, in figure 1, the x axis in the figure contains the list of all components, while the y axis shows the percentage of variance: each bar shows the Variance Ratio value for a component. We can observe that PC1 is the most important component, because the value of its variance, 23.6%, is the highest among all components. Components with low variance ratio values, like PC12 (1.26%), can be removed because their contribution is marginal. The curve in the graph describes how the variance ratio percentage varies when considering a variable number of components. From left to right, it ranges from the most important, to the least important component. The curve in figure 1 shows that in order to cover 95% of the variance ratio range, components from 10 to 12 are not required. Similarly, in order to achieve 85%, only 8 components are necessary: these 8 will be used for the clustering in our evaluation presented later in the paper.

TABLE IV
Experimental Configurations

Experiment	Algorithm	Normalization	Variable Set
Exp1	k-means	Min-Max	All
Exp2	k-medoids	Min-Max	All
Exp3	k-means	Categorized	All
Exp4	k-medoids	Categorized	All
Exp5	k-means	Min-Max	CC
Exp6	k-medoids	Min-Max	CC
Exp7	k-means	Standard	PCA_{95}
Exp8	k-means	Standard	PCA_{85}

4. Validation and Results

The goal of this work was to identify the *best* cluster of *good* customers so that future customers could be classified using this dataset. Thus, the goal of our validation is to measure and validate clusters of *good* customers, identify the best overall cluster and provide explainable evidence to support this selection.

In table IV, each row is a different experiment, where columns describe its configuration. Specifically for the columns: *Experiment* is a label to describe the experiment type; *Algorithm* is the clustering method used; *Normalization* is the type of normalization adopted; and, *Variable Set* defines the selection of variables used. Normalization can be of three types: *Min-Max*, where values for each attribute are on a (0,1) scale; *Categorized*, where all variables are transformed to a scale of 1-4 representing percentiles and Standard Scale as used by the PCA algorithm. For Variable Set we have the following options: *All*, indicates the full set of 12 variables; *CC* indicates the variable set recommended by the correlation coefficient method; and *PCA* indicates the variable set generated by PCA with a variance within 95% or 85%.

4.1. Validation Methodology

In this section, we present the strategy we adopt to evaluate the clusters obtained from the different experiment configurations. Specifically, we implemented two validation methods: a Query-based Validation (QA), and the a Recency-Frequency-Monetary validation (RFM). The RFM (Christy et al. (2018)) method is a well known approach, while the QA method is a customized approach resulting from discussions with our industry partners. We now provide an understanding of how each validation method works and the thresholds that can be applied for varying the results.

4.1.1. RFM Validation

RFM validation is a very popular model for classifying customer value. The RFM method is based on **R**ecency (R): how recent was the last purchase; **F**requency (F): how often do they purchase a policy; and **M**onetary Value (M): how much they spent. In order to calculate a customer score, these metrics must be associated with values. Metric values

range from 1 to 4, where 1 is the best score and 4 the worst. The function defining the values for R is described in definition 1.

Definition 1. *Given a customer* ***c****, the year* ***d*** *of the last purchase or renewal of a policy, and the year D of the most recent policy in the dataset, then the score $\mathcal{R}$ for recency is:*

$$\mathcal{R}(c,d,D) = \begin{cases} 1, & \textit{if } d = D \\ 2, & \textit{if } d = D-1 \\ 3, & \textit{if } d = D-2 \\ 4, & \textit{if } d <= D-3 \end{cases}$$

For example, the most recent policy was purchased (renewed) in 2015, therefore the value for R are: 1 for policies from 2015, 2 for policies from 2014, 3, for policies from 2013, and 4 for all others. F and M are described by variables YNbr and Margin. Values for F and M are defined by the percentile brackets the values for YNbr and Margin in table III fall into, as per definition 2.

Definition 2. *Given a customer* ***c****, and* ***v*** *the value from table II, then score $\mathcal{F}$ for frequency and the score $\mathcal{M}$ for monetary value are defined as, respectively:*

$$\mathcal{F}(c,v), \mathcal{M}(c,v) = \begin{cases} 1, & \textit{if } v \textit{ is equal or higher than the } 75^{th} \textit{ percentile} \\ 2, & \textit{if } v \textit{ is between the } 75^{th} \textit{ and the } 50^{th} \textit{ percentile} \\ 3, & \textit{if } v \textit{ is between the } 50^{th} \textit{ and the } 25^{th} \textit{ percentile} \\ 4, & \textit{Otherwise} \end{cases}$$

In simple terms, recency is lower if the policy purchase or renewal is more recent while frequency and monetary values are lower if the attributes are in the highest percentile. For RFM, we define a *good* customer as having a score of 1 across *all* metrics.

4.1.2. Query-Based Validation

For the QA (query-based) validation we define five different criteria, or queries. Queries are mutually independent, each measuring a different aspect of the customer (compared to the cluster). Each query is defined on a single variable from section 3. The general definition of a query is presented in definition 3.

Definition 3. *Given a customer* ***c****, the function $\mathcal{A}_v$ defining the value of a set of customers given a variable* ***v****, and a cluster $\mathcal{C}$ of customers: a query-validation $\mathcal{VD}$ is defined as:*

$$\mathcal{VD}(c,v,C) = \mathcal{A}_v(c) > Average(\mathcal{A}_v(C)) \tag{2}$$

where $\boldsymbol{v} \in \boldsymbol{V}$*, and with* $V = \{Margin, CCNber, CFNber, Maxh, YNbr\}$

The query validation in definition 3 generates one query for each variable in V. Each query defines whether a customer is *good* or not. A good customer is defined as one having value above the average. The variables considered for these queries are *Margin, CCNber, CFNber, Maxh,* and *YNbr*, with table V showing the query validation set after instantiating definition 3 for each variable. These were selected because:

TABLE V
Query Validation Set for selected variables

Name	Query
VQ1	$\mathcal{VD}(c, Margin) = \mathcal{A}_v(c) > Average(\mathcal{A}_{Margin}(C))$
VQ2	$\mathcal{VD}(c, CCNber) = \mathcal{A}_v(c) > Average(\mathcal{A}_{CCNber}(C))$
VQ3	$\mathcal{VD}(c, CFNber) = \mathcal{A}_v(c) > Average(\mathcal{A}_{CFNber}(C))$
VQ4	$\mathcal{VD}(c, Maxh) = \mathcal{A}_v(c) > Average(\mathcal{A}_{Maxh}(C))$
VQ5	$\mathcal{VD}(c, YNber) = \mathcal{A}_v(c) > Average(\mathcal{A}_{YNber}(C))$

- Margin, and YNbr represent the customer margin and the number of years a customer stayed with the company, and are two metrics widely used in the literature, e.g. Sohrabi and Khanlari (2007); Khajvand et al. (2011);
- CCNber is related to Margin and YNbr, and reflects the number of policies a customer holds;
- CFNber reflects that family member policies contribute to the customer;
- Maxh is the maximum number of policies the customer had during a year.

4.2. Results

Before presenting our results, it is useful to restate the goals of our research. The primary goal is to determine if customers could be grouped according to their *value*. It is easier for managers and analysts to understand these groupings if there are only a small number of groups. As clusters increase in number, it can be difficult to distinguish them. Secondly, market analysts want to plan budget spend and thus, want to identify the *good* customers. Using table IV, the evaluation has 8 distinct experiments but within each experiment, every clustering algorithm has values for k from 1 to 5, meaning 5 experimental runs within each experiment. The result was 15 cluster datasets generated during *every* experiment as can be seen in table VI. As 120 clusters (8x15) represents too much information for discussion, we selected the best performing experiments and in table VI, present the clusters and validation results for discussion.

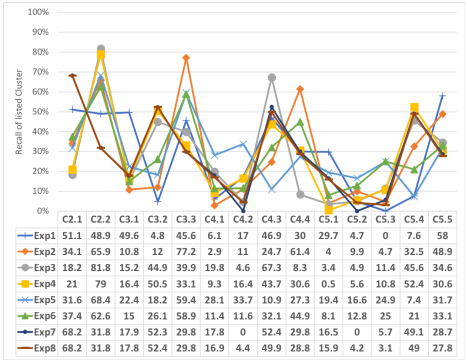
In section 4.2.1, we discuss our use of recall and precision to identify the best performing experiment and in section 4.2.2, we highlight the best performing set of clusters within that experiment. In the final part of this section, we discuss the implications of these results with respect to decision making for managers and analysts.

4.2.1. Recall and Precision Results

In this section, we present an overview of the results for 8 outlined in table IV, using evaluation metrics VQ1 to VQ5 and RFM together with recall and precision to identify the best performing experimental configuration.

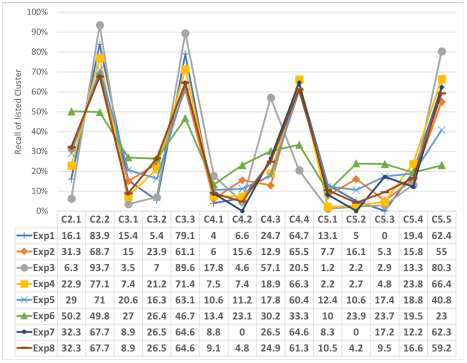
Assume that TC represents the total number of (customer) objects classified as *good* in the entire dataset; TG the number of objects in a cluster; and CG the number of

customers classified as good in that cluster. Both validations are equally important: *Recall* represents the fraction of *good* customers in the cluster, divided by the overall number of good customers $\frac{CG}{TC}$; and *Precision* is the fraction of *good* customers in a cluster $\frac{CG}{TG}$.



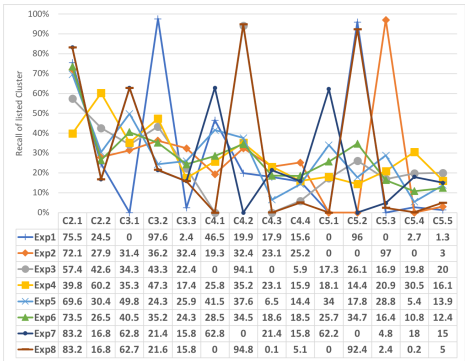
	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	51.1	48.9	49.6	4.8	45.6	6.1	17	46.9	30	29.7	4.7	0	7.6	58
Exp2	34.1	65.9	10.8	12	77.2	2.9	11	24.7	61.4	4	9.9	4.7	32.5	48.9
Exp3	18.2	81.8	15.2	44.9	39.9	19.8	4.6	67.3	8.3	3.4	4.9	11.4	45.6	34.6
Exp4	21	79	16.4	50.5	33.1	9.3	16.4	43.7	30.6	0.5	5.6	10.8	52.4	30.6
Exp5	31.6	68.4	22.4	18.2	59.4	28.1	33.7	10.9	27.3	19.4	16.6	24.9	7.4	31.7
Exp6	37.4	62.6	15	26.1	58.9	11.4	11.6	32.1	44.9	8.1	12.8	25	21	33.1
Exp7	68.2	31.8	17.9	52.3	29.8	17.8	0	52.4	29.8	16.5	0	5.7	49.1	28.7
Exp8	68.2	31.8	17.8	52.4	29.8	16.9	4.4	49.9	28.8	15.9	4.2	3.1	49	27.8

(A) Recall VQ1 for all experiment



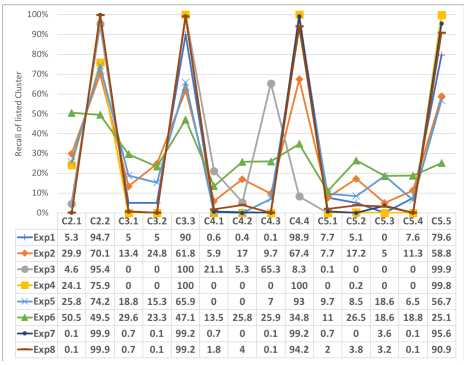
	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	16.1	83.9	15.4	5.4	79.1	4	6.6	24.7	64.7	13.1	5	0	19.4	62.4
Exp2	31.3	68.7	15	23.9	61.1	6	15.6	12.9	65.5	7.7	16.1	5.3	15.8	55
Exp3	6.3	93.7	3.5	7	89.6	17.8	4.6	57.1	20.5	1.2	2.2	2.9	13.3	80.3
Exp4	22.9	77.1	7.4	21.2	71.4	7.5	7.4	18.9	66.3	2.2	2.7	4.8	23.8	66.4
Exp5	29	71	20.6	16.3	63.1	10.6	11.2	17.8	60.4	12.4	10.6	17.4	18.8	40.8
Exp6	50.2	49.8	27	26.4	46.7	13.4	23.1	30.2	33.3	10	23.9	23.7	19.5	23
Exp7	32.3	67.7	8.9	26.5	64.6	8.8	0	26.5	64.6	8.3	0	17.2	12.2	62.3
Exp8	32.3	67.7	8.9	26.5	64.6	9.1	4.8	24.9	61.3	10.5	4.2	9.5	16.6	59.2

(B) Recall VQ2 for all experiment



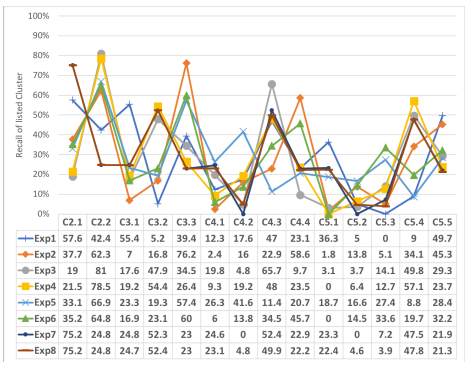
	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	75.5	24.5	0	97.6	2.4	46.5	19.9	17.9	15.6	0	96	0	2.7	1.3
Exp2	72.1	27.9	31.4	36.2	32.4	19.3	32.4	23.1	25.2	0	0	97	0	3
Exp3	57.4	42.6	34.3	43.3	22.4	0	94.1	0	5.9	17.3	26.1	16.9	19.8	20
Exp4	39.8	60.2	35.3	47.3	17.4	25.8	35.2	23.1	15.9	18.1	14.4	20.9	30.5	16.1
Exp5	69.6	30.4	49.8	24.3	25.9	41.5	37.6	6.5	14.4	34	17.8	28.8	5.4	13.9
Exp6	73.5	26.5	40.5	35.2	24.3	28.5	34.5	18.6	18.5	25.7	34.7	16.4	10.8	12.4
Exp7	83.2	16.8	62.8	21.4	15.8	62.8	0	21.4	15.8	62.2	0	4.8	18	15
Exp8	83.2	16.8	62.7	21.6	15.8	0	94.8	0.1	5.1	0	92.4	2.4	0.2	5

(C) Recall VQ3 for all experiment



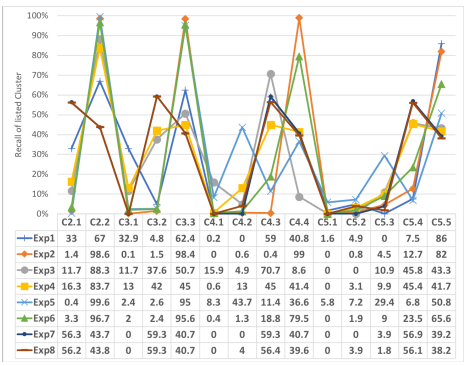
	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	5.3	94.7	5	5	90	0.6	0.4	0.1	98.9	7.7	5.1	0	7.6	79.6
Exp2	29.9	70.1	13.4	24.8	61.8	5.9	17	9.7	67.4	7.7	17.2	5	11.3	58.8
Exp3	4.6	95.4	0	0	100	21.1	5.3	65.3	8.3	0.1	0	0	0	99.9
Exp4	24.1	75.9	0	0	100	0	0	0	100	0	0.2	0	0	99.8
Exp5	25.8	74.2	18.8	15.3	65.9	0	0	7	93	9.7	8.5	18.6	6.5	56.7
Exp6	50.5	49.5	29.6	23.3	47.1	13.5	25.8	25.9	34.8	11	26.5	18.6	18.8	25.1
Exp7	0.1	99.9	0.7	0.1	99.2	0.7	0	0.1	99.2	0.7	0	3.6	0.1	95.6
Exp8	0.1	99.9	0.7	0.1	99.2	1.8	4	0.1	94.2	2	3.8	3.2	0.1	90.9

(D) Recall VQ4 for all experiment



	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	57.6	42.4	55.4	5.2	39.4	12.3	17.6	47	23.1	36.3	5	0	9	49.7
Exp2	37.7	62.3	7	16.8	76.2	2.4	16	22.9	58.6	1.8	13.8	5.1	34.1	45.3
Exp3	19	81	17.6	47.9	34.5	19.8	4.8	65.7	9.7	3.1	3.7	14.1	49.8	29.3
Exp4	21.5	78.5	19.2	54.4	26.4	9.3	19.2	48	23.5	0	6.4	12.7	57.1	23.7
Exp5	33.1	66.9	23.3	19.3	57.4	26.3	41.6	11.4	20.7	18.7	16.6	27.4	8.8	28.4
Exp6	35.2	64.8	16.9	23.1	60	6	13.8	34.5	45.7	0	14.5	33.6	19.7	32.2
Exp7	75.2	24.8	24.8	52.3	23	24.6	0	52.4	22.9	23.3	0	7.2	47.5	21.9
Exp8	75.2	24.8	24.7	52.4	23	23.1	4.8	49.9	22.2	22.4	4.6	3.9	47.8	21.3

(E) Recall VQ5 for all experiment

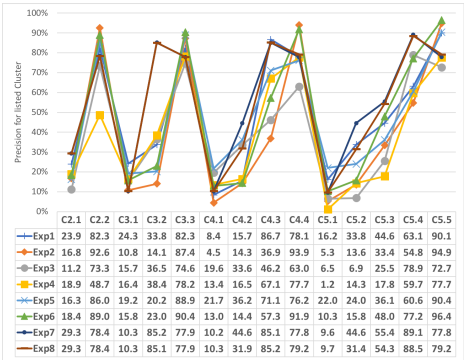


	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	33	67	32.9	4.8	62.4	0.2	0	59	40.8	1.6	4.9	0	7.5	86
Exp2	1.4	98.6	0.1	1.5	98.4	0	0.6	0.4	99	0	0.8	4.5	12.7	82
Exp3	11.7	88.3	11.7	37.6	50.7	15.9	4.9	70.7	8.6	0	0	10.9	45.8	43.3
Exp4	16.3	83.7	13	42	45	0.6	13	45	41.4	0	3.1	9.9	45.4	41.7
Exp5	0.4	99.6	2.4	2.6	95	8.3	43.7	11.4	36.6	5.8	7.2	29.4	6.8	50.8
Exp6	3.3	96.7	2	2.4	95.6	0.4	1.3	18.8	79.5	0	1.9	9	23.5	65.6
Exp7	56.3	43.7	0	59.3	40.7	0	0	59.3	40.7	0	0	3.9	56.9	39.2
Exp8	56.2	43.8	0	59.3	40.7	0	4	56.4	39.6	0	3.9	1.8	56.1	38.2

(F) Recall RFM for all experiment

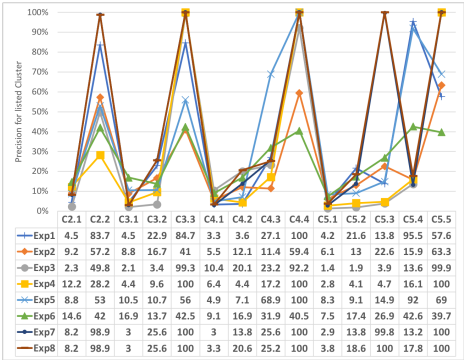
FIGURE 2: Recall Results (All Experiments)

Figure 2 presents the results for recall and figure 3 presents the result for precision. In each figure, we have 6 graphs. In each graph, each VQ and RFM validation result for all experiments, are presented for comparison. The results in the graph are the percentage of either recall or precision in a specified cluster, with the results kept to 1 decimal place.



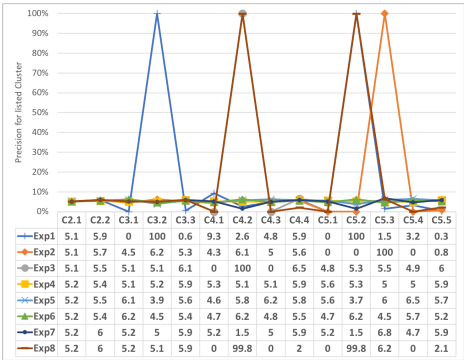
	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	23.9	82.3	24.3	33.8	82.3	8.4	15.7	86.7	78.1	16.2	33.8	44.6	63.1	90.1
Exp2	16.8	92.6	10.8	14.1	87.4	4.5	14.3	36.9	93.9	5.3	13.6	33.4	54.8	94.9
Exp3	11.2	73.3	15.7	36.5	74.6	19.6	33.6	46.2	63.0	6.5	6.9	25.5	78.9	72.7
Exp4	18.9	48.7	16.4	38.4	78.2	13.4	16.5	67.1	77.7	1.2	14.3	17.8	59.7	77.7
Exp5	16.3	86.0	19.2	20.2	88.9	21.7	36.2	71.1	76.2	22.0	24.0	36.1	60.6	90.4
Exp6	18.4	89.0	15.8	23.0	90.4	13.0	14.4	57.3	91.9	10.3	15.8	48.0	77.2	96.4
Exp7	29.3	78.4	10.3	85.2	77.9	10.2	44.6	85.1	77.8	9.6	44.6	55.4	89.1	77.8
Exp8	29.3	78.4	10.3	85.1	77.9	10.3	31.9	85.2	79.2	9.7	31.4	54.3	88.5	79.2

(A) Precision of VQ1 for all experiment



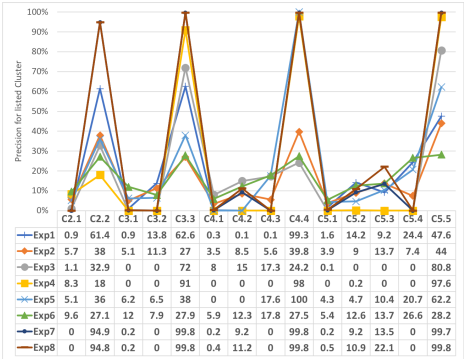
	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	4.5	83.7	4.5	22.9	84.7	3.3	3.6	27.1	100	4.2	21.6	13.8	95.5	57.6
Exp2	9.2	57.2	8.8	16.7	41	5.5	12.1	11.4	59.4	6.1	13	22.6	15.9	63.3
Exp3	2.3	49.8	2.1	3.4	99.3	10.4	20.1	23.2	92.2	1.4	1.9	3.9	13.6	99.9
Exp4	12.2	28.2	4.4	9.6	100	6.4	4.4	17.2	100	2.8	4.1	4.7	16.1	100
Exp5	8.8	53	10.5	10.7	56	4.9	7.1	68.9	100	8.3	9.1	14.9	92	69
Exp6	14.6	42	16.9	13.7	42.5	9.1	16.9	31.9	40.5	7.5	17.4	26.9	42.6	39.7
Exp7	8.2	98.9	3	25.6	100	3	13.8	25.6	100	2.9	13.8	99.8	13.2	100
Exp8	8.2	98.9	3	25.6	100	3.3	20.6	25.2	100	3.8	18.6	100	17.8	100

(B) Precision of VQ2 for all experiment



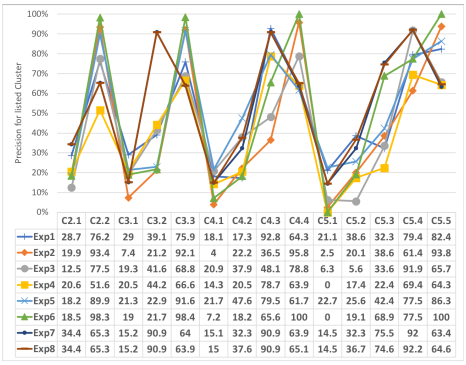
	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	5.1	5.9	0	100	0.6	9.3	2.6	4.8	5.9	0	100	1.5	3.2	0.3
Exp2	5.1	5.7	4.5	6.2	5.3	4.3	6.1	5	5.6	0	0	100	0	0.8
Exp3	5.1	5.5	5.1	5.1	6.1	0	100	0	6.5	4.8	5.3	5.5	4.9	6
Exp4	5.2	5.4	5.1	5.2	5.9	5.3	5.1	5.1	5.9	5.6	5.3	5	5	5.9
Exp5	5.2	5.5	6.1	3.9	5.6	4.6	5.8	6.2	5.8	5.6	3.7	6	6.5	5.7
Exp6	5.2	5.4	6.2	4.5	5.4	4.7	6.2	4.8	5.5	4.7	6.2	4.5	5.7	5.2
Exp7	5.2	6	5.2	5	5.9	5.2	1.5	5	5.9	5.2	1.5	6.8	4.7	5.9
Exp8	5.2	6	5.2	5.1	5.9	0	99.8	0	2	0	99.8	6.2	0	2.1

(C) Precision of VQ3 for all experiment



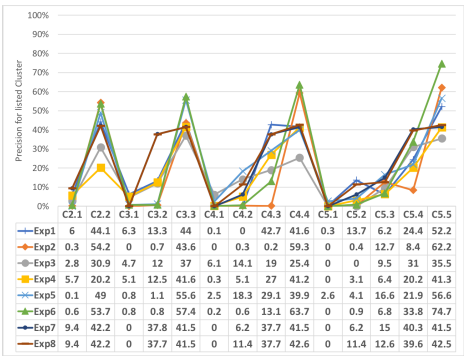
	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	0.9	61.4	0.9	13.8	62.6	0.3	0.1	0.1	99.3	1.6	14.2	9.2	24.4	47.6
Exp2	5.7	38	5.1	11.3	27	3.5	8.5	5.6	39.8	3.9	9	13.7	7.4	44
Exp3	1.1	32.9	0	0	72	8	15	17.3	24.2	0.1	0	0	0	80.8
Exp4	8.3	18	0	0	91	0	0	0	98	0	0.2	0	0	97.6
Exp5	5.1	36	6.2	6.5	38	0	0	17.6	100	4.3	4.7	10.4	20.7	62.2
Exp6	9.6	27.1	12	7.9	27.9	5.9	12.3	17.8	27.5	5.4	12.6	13.7	26.6	28.2
Exp7	0	94.9	0.2	0	99.8	0.2	9.2	0	99.8	0.2	9.2	13.5	0	99.7
Exp8	0	94.8	0.2	0	99.8	0.4	11.2	0	99.8	0.5	10.9	22.1	0	99.8

(D) Precision of VQ4 for all experiment



	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	28.7	76.2	29	39.1	75.9	18.1	17.3	92.8	64.3	21.1	38.6	32.3	79.4	82.4
Exp2	19.9	93.4	7.4	21.2	92.1	4	22.2	36.5	95.8	2.5	20.1	38.6	61.4	93.8
Exp3	12.5	77.5	19.3	41.6	68.8	20.9	37.9	48.1	78.8	6.3	5.6	33.6	91.9	65.7
Exp4	20.6	51.6	20.5	44.2	66.6	14.3	20.5	78.7	63.9	0	17.4	22.4	69.4	64.3
Exp5	18.2	89.9	21.3	22.9	91.6	21.7	47.6	79.5	61.7	22.7	25.6	42.4	77.5	86.3
Exp6	18.5	98.3	19	21.7	98.4	7.2	18.2	65.6	100	0	19.1	68.9	77.5	100
Exp7	34.4	65.3	15.2	90.9	64	15.1	32.3	90.9	63.9	14.5	32.3	75.5	92	63.4
Exp8	34.4	65.3	15.2	90.9	63.9	15	37.6	90.9	65.1	14.5	36.7	74.6	92.2	64.6

(E) Precision of VQ5 for all experiment



	C2.1	C2.2	C3.1	C3.2	C3.3	C4.1	C4.2	C4.3	C4.4	C5.1	C5.2	C5.3	C5.4	C5.5
Exp1	6	44.1	6.3	13.3	44	0.1	0	42.7	41.6	0.3	13.7	6.2	24.4	52.2
Exp2	0.3	54.2	0	0.7	43.6	0	0.3	0.2	59.3	0	0.4	12.7	8.4	62.2
Exp3	2.8	30.9	4.7	12	37	6.1	14.1	19	25.4	0	0	9.5	31	35.5
Exp4	5.7	20.2	5.1	12.5	41.6	0.3	5.1	27	41.2	0	3.1	6.4	20.2	41.3
Exp5	0.1	49	0.8	1.1	55.6	2.5	18.3	29.1	39.9	2.6	4.1	16.6	21.9	56.6
Exp6	0.6	53.7	0.8	0.8	57.4	0.2	0.6	13.1	63.7	0	0.9	6.8	33.8	74.7
Exp7	9.4	42.2	0	37.8	41.5	0	6.2	37.7	41.5	0	6.2	15	40.3	41.5
Exp8	9.4	42.2	0	37.7	41.5	0	11.4	37.7	42.6	0	11.4	12.6	39.6	42.5

(F) Precision of RFM for all experiment

FIGURE 3: Precision Results (All Experiments)

Within each experiment, clusters are named as Ckn, where k is the *cluster count* parameter for the clustering algorithm and n the number of the cluster within that experiment, e.g. C3.2 indicates the 2nd cluster in a 3-cluster (k=3) experiment.

For each k={2,3,4,5}, the recall result for the aggregate will always be 100% because recall calculates the fraction of *good* from the overall total of *good* customers. However, for precision results, we would like to see a very high value for a single cluster with remaining clusters in the same k-experiment having low values. This suggests that *good*

customers are assigned (more or less) to the same high scoring cluster.

It is possible for the top ranked cluster to appear more than once in a validation. For example, experiment Exp1 ranks C3.2 as best in VQ3 for both recall *and* precision. Moreover, the customers for cluster C4.4 in Exp1 are all (100%) *good* customers using VQ2 and clusters C3.2 and C5.2 are all (100%) *good* customers using VQ3. Using the 2 best clusters in each experiment, we applied a score of 2 for the top performing cluster and score of 1 for the next best, when generating rankings.

- For *recall*: Experiment Exp3 ranks highest with 4 top ranked matches and 1 second place with a score of 9; Exp4 is next best with a total score of 6; with Exp2 3rd best with a score of 3.
- Exp6 ranks highest for *precision* with a total score of 6; Exp2 is next best with a score of 5; Exp3 scored 3 for *precision*.
- If we simply count results for both recall and precision, experiment 3 performs best as it has the highest number of top (5) and second ranked (2) clusters, with a score of 12 using our simple metric. Thus, experiment 3 is used for our discussion in the following section.

4.2.2. Interpreting the best Configuration

Table VI presents data from the best ranking experiment, as selected by the recall and precision results, with a configuration (from table IV) of {k-means, Categorized, All}. The goal of this part of the evaluation is to determine which k-cluster configuration for experiment 3 out-performs others by best classifying *good* customers into a single cluster.

The column **Entities** lists the number of objects assigned to this cluster. The columns **Margin**, **CCNbr**, **CFNbr**, **Maxh**, and **YNbr** show the average of the listed variables in a certain cluster. The columns **VQ1**, **VQ2**, **VQ3**, **VQ4**, **VQ5**, and **RFM** present the number of matches in the cluster by the validation methods. The higher the number, the more *good* customers are found in this cluster. The **Union** is listed for the objects that fulfill any 4 validations in {VQ1 to VQ5 and RFM}. Finally, the column AF-Score identifies the most desirable cluster by summing recall and precision scores.

Using recall and precision scores, we now introduce the algorithm *AF-Score*, which is a measure of the *desirability* of a cluster: the higher the AF-Score for a cluster, the better the customers in that cluster are classified (good customers). The AF-Score is the sum of F-score values for all 6 validation rules, applied to each cluster and formally presented in definition 4.

Definition 4. *Given a cluster C, and a validation rule VR$_i$, and $recall_i$, and $precision_i$ representing the recall and precision for C and VR$_i$ respectively, the AF-Score is:*

$$\textit{AF-Score}(C) = \sum_{i=0}^{|VRS|} 2 * \frac{recall_i * precision_i}{recall_i + precision_i}$$

where VR$_i \in$ VRS, with VRS $= \{$VQ1, VQ2, VQ3, VQ4, VQ5, RFM $\}$

TABLE VI
Experiment 3 Results Across k Configurations

Clu.	Entities	Margin	CCNbr	CFNbr	Maxh	YNbr	VQ1	VQ2	VQ3	VQ4	VQ5	RFM	Union	AF-Score
C1	387951	1336.40	1.29	0.05	1.16	2.57	141783	84071	20468	54626	151336	55439	50080	205.0%
C2.1	229702	790.78	1.02	0.05	1.01	1.73	25841	5276	11743	2521	28764	6476	353	48.1%
C2.2	158249	2129.59	1.68	0.06	1.37	3.79	115942	78795	8725	52105	122572	48963	49727	326.0%
C3.1	137876	862.06	1.02	0.05	1.00	1.92	21598	2932	7011	0	26574	6476	805	52.0%
C3.2	174236	1287.14	1.04	0.05	1.00	2.60	63593	5844	8868	0	72547	20865	2742	116.6%
C3.3	75839	2314.45	2.37	0.06	1.80	3.67	56592	75295	4589	54626	52215	28098	46533	328.2%
C4.1	143427	947.93	1.12	0.00	1.09	2.00	28054	14952	0	11545	29934	8799	4754	73.6%
C4.2	19253	1286.91	1.28	1.00	1.17	2.54	6477	3876	19253	2884	7291	2707	3994	136.1%
C4.3	206577	1572.61	1.32	0.00	1.19	2.88	95480	48006	0	35670	99389	39180	31156	200.6%
C4.4	18694	1767.86	2.35	0.07	1.27	3.57	11772	17237	1215	4527	14722	4753	10176	96.9%
C5.1	73950	657.33	1.01	0.05	1.00	1.36	4810	1016	3532	52	4665	0	79	17.4%
C5.2	101169	733.39	1.02	0.05	1.00	1.55	7005	1872	5337	0	5635	0	0	21.0%
C5.3	63259	1088.99	1.04	0.05	1.00	2.55	16143	2479	3462	0	21264	6023	1171	57.3%
C5.4	81990	2090.08	1.15	0.05	1.00	4.14	64723	11183	4051	0	75387	25416	9046	180.8%
C5.5	67583	2302.20	2.40	0.06	1.89	3.54	49102	67521	4086	54574	44385	24000	39784	314.1%

When focusing on clusters of good customers (C2.2, C3.3, C5.5), their average for most variables is highest. Management can interpret that customers in these clusters appear to stay longer, buy more policies, and pay more money to the company. As a result, the evaluation selects these 3 candidate clusters. We can also look at some of the highlights presented in table VI.

It is shown that for the total number of 387,951 customers there are 141,783 (VQ1) clients with a margin greater than 1336.41 euros; 84,071 (VQ2) customers (CCNber) with more than 1 policy; 20,468 (VQ3) customers with a family member (CFNber) who also holds a policy; 54,629 (VQ4) customers who have held more than 1 policy with the company; and 151,336 (VQ5) customers who stayed more than 2 years. This is the data which drives the clustering strategy.

In total, there are 55,439 *good* clients classified using the RFM method, where R=F=M=1 and there are 50,080 customers who meet at least 4 validation criteria (the Union column). For any k={2,3,4,5}, clusters C2.2, C3.3, and C5.5 contained more than 80% of *good* customers (Perc) with a maximum of 41% assigned to cluster C2.2 ($\frac{158249}{387951}$=41%). Cluster C3.3 is noteworthy with only 20% of customers but 93% of matches, indicating a small but very accurate cluster, meaning managers can have a high degree of confidence with this cluster.

Apart from highlighting the details within each high performing cluster, it is also worth examining the clusters at a higher level. Where managers might opt for a larger number of customer clusters and the value of k increases, good customers are spread across multiple clusters. For k = 2, over 80% of good customers lie in cluster C2.2 using 5 validation criteria. For k=3, most were assigned to clusters C3.2 and C3.3. For k=4, some good customers were placed into cluster C4.3, and for k=5, a number were assigned to C5.4. The guidance here is that where managers have little knowledge of the dataset, set a low value for k to ensure that the majority of good customers can be located in a single cluster.

In general, C3.3 outperforms others using AF-Score values, meaning it is the best cluster in terms of good customers. However, the performance of C2.1 and C5.5 clusters are also candidates for management seeking the optimum grouping of good customers.

For those clusters who performed well in AF-Score, most of the time, both recall and precision for a validation method are higher than other clusters in this set of k. A high AF-Score generally means both recall and precision are high. However, while cluster C4.4 has a high precision value, the number of objects in this cluster (TG) is low and as a result, this decreases the recall value, which has the effect of reducing the AF-Score. This provides a level of uncertainty for managers and perhaps a sign that these clusters should be avoided as budget spend may not be maximized.

The *CFNbr* variable (having a family member) represents an interesting outlier. For k=4, C4.2 clustered more than 94% of customers who had a family member. For other k experiments, customers with *CFNbr* values are spread into different clusters. There is very little research in the literature where family members are used to support decision making and we believe this to be an area that could benefit from further research.

Variable selection is an important process for prediction algorithms and we employed two of the more popular methods. The result from our evaluation provides strong evidence that configurations using *all* variables outperform all other variable combinations (Exp3 performs best while Exp2 is second best) when ranking the best clusters. A clear message emerges from these experiments: when calculating CLV values, organizations generally do not have a large variable set and only large variable sets benefit from variable selection methods. This is true even where variables are highly correlated. Where the results from precision and recall experiments are very similar, e.g. VQ1 and VQ5, there is a strong possibility that this is caused by two highly correlated variables. This insight informs management on what variables and how many, to extract from operational databases.

However, the datasets generated using Correlation Coefficient (CC) and PCA should not be discarded completely. While both CC and PCA variable sets perform badly in Recall experiments, they perform well in Precision experiments: the CC variable set outperforms others, having the highest (or second highest) accuracy at VQ2, VQ4, VQ5 and RFM and the PCA variable set performs well in VQ2, VQ3, VQ4. This is because CC and PCA variables are very accurate at identifying good customers. The issue is that the size of the good customer cluster is small. However, this does provide an interesting insight for follow-on experiments that use these variable selection techniques.

4.3. Summary and Managerial Implications

A key requirement for managers is to segment customers according to future spending and thus, our main question is if we could successfully auto-classify the customer dataset. The results of a robust evaluation indicate that this goal has been achieved. Our method successfully classifies customer datasets with an accuracy of 90%, as shown by the results in table VI. It is important to point out that while we can accurately classify *good* customers, if management requires more granular groupings (as k increases), this is more difficult. However, our experiments highlight that by examining the *average value* for the key variables Margin, CCNbr, CFNbr, Maxh, and YNbr in each cluster (table VI), an algorithm can label the group of clusters with an accuracy of 99.3%. This is a finding that we did not anticipate prior to performing our experiments and informs management of the importance of collecting or calculating these values.

From a management perspective, the determination of each customer's value to the organization provides significant support in future budgeting decisions. However, this remains an open problem for organizations with either the simple approaches generally taken, or the highly theoretical approaches offered in the literature. While the ultimate goal is a personalized CLV value for every customer, there are steps to achieving this that provide benefits to managers and marketing analysts. If customer lifetime values remain outside the reach of managers, this implies that all of their datasets are untrained or key variables are unavailable. The impact of our work is to support managers and analysts by providing these broad classifications of customers with an understanding as to *how* these decisions were reached. Table IV informs management as to which variables provide the best results: these variables or the data used to compute them should become mandatory inputs on future transactions.

Section 4.2.2 describes the process by which the best performing experimental configuration was chosen. This may suit the needs of management with a very specific requirement. However, as we expose the decision making process and the metrics involved, managers may decide on different experimental configurations or may select the results from more than one experiment to increase the number of *good* customers. They may also decide to use a different variable set based on the understanding we have provided about specific variables and their impact on defining clusters of good customers.

5. Conclusions and Future Plans

Many organizations interact with their customers through different platforms with the insurance industry providing a good example. New and renewed polices can be generated through different online portals, through different types of brokers or in a high street shop. As a result, a full customer history is rarely complete and calculating customer metrics such as Customer Lifetime Value is difficult or inaccurate. To go from a disconnected customer record to a dataset which supports decision support operations such as calculating Customer Lifetime Value requires: linking each customer record, classifying customers according to broad but specific categories, before imputing the final variables required for calculations of customer metrics.

We believe that this work makes a significant contribution as it automates the classification of customers. Our research focuses on the broad classification of customers after their full history has been compiled and attempts to provide management with an insight as to how these classifications are achieved. By providing this level of understanding, management can customize datasets or clustering parameters to segment customers according to specific needs.

A limitation to this work involves the feature set used. As we used a real-world dataset provided by an industry partner, this meant that not all the variables needed to classify customers were available or present. While this reflects a real and common problem for organizations, the two most popular variables for CLV calculations were missing: acquisition and retention. Thus, the imputation of these values are part of our future research and will be published in Nie et al. (2020).

References

Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L., and Weaven, S. (2019). Market segmentation and travel choice prediction in spa hotels through tripadvisor's online reviews. *International Journal of Hospitality Management*, 80:52–77.

Asuero, G. A., Sayago, A., and González, G. (2006). The correlation coefficient: An overview. *Critical Reviews in Analytical Chemistry*, 36(1):41–59.

Barnes, S. J., Bauer, H. H., Neumann, M. M., and Huber, F. (2007). Segmenting cyberspace: a customer typology for the internet. *European journal of marketing*, 41(1/2):71–93.

Berger, P. D. and Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*, 12(1):17 – 30.

Blattberg, R. C. and Deighton, J. (1996). Manage marketing by the customer equity test. *Harvard business review*, 74(4):136.

Christy, A. J., Umamakeswari, A., Priyatharsini, L., and Neyaa, A. (2018). Rfm ranking–an effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*.

Cibulková, J. and Sulc, Z. (2018). A case study of customer segmentation with the use of hierarchical cluster analysis of categorical data. In *Proceedings of the Applications of Mathematics and Statistics in Economics, AMSE, Kutná Hora, Czech Republic, August 29 – September 2, 2018*.

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, 1(3):131–156.

Di Benedetto, C. A. and Kim, K. H. (2016). Customer equity and value management of global brands: Bridging theory and practice from financial and marketing perspectives: Introduction to a journal of business research special section. *Journal of Business Research*, 69(9):3721–3724.

Fader, P., Hardie, B., and Berger, P. (2004). Customer-base analysis with discrete-time transaction data. *SSRN Electronic Journal*.

Fader, P., Hardie, B., and Lee, K. (2005). Rfm and clv: Using iso-value curves for customer base analysis. *Journal of Marketing Research American Marketing Association ISSN*, XLII:415–430.

Fotiadis, A. K. and Vassiliadis, C. (2017). Being customer-centric through CRM metrics in the B2B market: the case of maritime shipping. *Journal of Business and Industrial Marketing*, 32(3):347–356.

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., and Maggio, R. M. (2018). Use of principal component analysis (pca) and hierarchical cluster analysis (hca) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science & Technology*, 72:83–90.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. (2008). *Feature extraction: foundations and applications*, volume 207. Springer.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hillebrand, B. and Biemans, W. G. (2011). Dealing with downstream customers: an exploratory study. *Journal of Business & Industrial Marketing*, 26(2):72–80.

Hiziroglu, A. and Sengul, S. (2012). Investigating two customer lifetime value models from segmentation perspective. *Procedia-Social and Behavioral Sciences*, 62:766–774.

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.

Kara, A. and Kaynak, E. (1997). Markets of a single customer: exploiting conceptual developments in market segmentation. *European journal of marketing*, 31(11/12):873–895.

Keiningham, T., Perkins-Munn, T., Aksoy, L., and Estrin, D. (2005). Does customer satisfaction lead to profitability?: The mediating role of share-of-wallet. *Managing Service Quality*, 15:172–181.

Khajvand, M., Zolfaghar, K., Ashoori, S., and Alizadeh, S. (2011). Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 3:57 – 63. World Conference on Information Technology.

Kim, S.-Y., Jung, T.-S., Suh, E.-H., and Hwang, H.-S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert systems with applications*, 31(1):101–107.

King, J. R. and Jackson, D. A. (1999). Variable selection in large environmental data sets using principal components analysis. *Environmetrics*, 10(1):67–77.

Kleinaltenkamp, M., Rudolph, M., and Classen, M. (2012). Multistage marketing. In *Business-to-Business Marketing Management: Strategies, Cases, and Solutions*. Emerald Group Publishing Limited.

Lawrence, I. and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268.

Lee, P. H. and Philip, L. (2010). Distance-based tree models for ranking data. *Computational Statistics & Data Analysis*, 54(6):1672–1682.

Ling, R. and Yen, D. C. (2001). Customer relationship management: An analysis framework and implementation strategies. *Journal of computer information systems*, 41(3):82–97.

Lu, Z., Peiyi, W., Ping, C., Xianglong, L., Baoqun, Z., and Longfei, M. (2019). Customer segmentation algorithm based on data mining for electric vehicles. In *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 77–83. IEEE.

Malthouse, E. C. and Blattberg, R. C. (2005). Can we predict customer lifetime value? *Journal of Interactive Marketing*, 19(1):2 – 16.

Martínez-López, F. J. and Casillas, J. (2013). Artificial intelligence-based systems applied in industrial marketing: An historical overview, current and future insights. *Industrial Marketing Management*, 42(4):489 – 495. Special Issue on Applied Intelligent Systems in Business-to-Business Marketing.

Merrilees, B., Bentley, R., and Cameron, R. (1999). Business service market segmentation: The case of electrical and mechanical building maintenance services. *Journal of Business and Industrial Marketing*, 14(2):151–161.

Milligan, G. W. and Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of classification*, 5(2):181–204.

Müllensiefen, D., Hennig, C., and Howells, H. (2018). Using clustering of rankings to explain brand preferences with personality and socio-demographic variables. *Journal of Applied Statistics*, 45(6):1009–1029.

Ngai, E., Xiu, L., and Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2, Part 2):2592 – 2602.

Nie, D. and Roantree, M. (2019). Detecting multi-relationship links in sparse datasets. In *ICEIS 2019, Volume 1*, pages 149–157. SciTePress.

Nie, D., Scriney, M., and Roantree, M. (2020). Predicting customer churn for insurance data. In *To Appear in The 22nd International Conference on Big Data Analytics and Knowledge Discovery - DaWaK2020*.

Noorizadeh, A., Mahdiloo, M., and Saen, R. F. (2013). Evaluating relative value of customers via data envelopment analysis. *Journal of Business and Industrial Marketing*, 28(7):577–588.

Parkhimenka, U., Tatur, M., and Khandogina, O. (2017). Unsupervised ranking of clients: machine learning approach to define a" good customer".

Paschen, J., Kietzmann, J., and Kietzmann, T. C. (2019). Artificial intelligence (AI) and its implications for market knowledge in B2B marketing. *Journal of Business and Industrial Marketing*, 34(7):1410–1419.

Pfeifer, P. E. (2005). The optimal ratio of acquisition and retention costs. *Journal of Targeting, Measurement and Analysis for Marketing*, 13(2):179–188.

Powers, T. L. and Sterling, J. U. (2008). Segmenting business-to-business markets: A micro-macro linking methodology. *Journal of Business and Industrial Marketing*, 23(3):170–177.

Reinartz, W. J. and Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of marketing*, 64(4):17–35.

Rezankova, H., Loster, T., and Husek, D. (2011). Evaluation of categorical data clustering. In *Advances in Intelligent Web Mastering–3*, pages 173–182. Springer.

Roantree, M. and Liu, J. (2014). A heuristic approach to selecting views for materialization. *Softw. Pract. Exp.*, 44(10):1157–1179.

Sánchez-Hernández, G., Chiclana, F., Agell, N., and Aguado, J. C. (2013). Ranking and selection of unsupervised learning marketing segmentation. *Knowledge-based systems*, 44:20–33.

Scriney, M., McCarthy, S., McCarren, A., Cappellari, P., and Roantree, M. (2019). Automating data mart construction from semi-structured data sources. *Comput. J.*, 62(3):394–413.

Shabanpour, H., Yousefi, S., and Saen, R. F. (2017). Forecasting efficiency of green suppliers by dynamic data envelopment analysis and artificial neural networks. *Journal of Cleaner Production*, 142:1098 – 1107. Special Volume on Improving natural resource management and human health to ensure sustainable societal development based upon insights gained from working within 'Big Data Environments'.

Singh, J., Flaherty, K., Sohi, R. S., Deeter-Schmelz, D., Habel, J., Meunier-FitzHugh, K. L., Malshe, A., Mullins, R., and Onyemah, V. (2019). Sales profession and professionals in the age of digitization and artificial intelligence technologies: concepts, priorities, and questions. *Journal of Personal Selling & Sales Management*, 39(1):2–22.

Smith, K. A., Willis, R. J., and Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: A case study. *Journal of the operational research society*, 51(5):532–541.

Sohrabi, B. and Khanlari, A. (2007). Customer lifetime value (clv) measurement based on rfm model. *Iranian Accounting & Auditing Review*, 14(47):7–20.

Sudharshan, D. and Winter, F. (1998). Strategic segmentation of industrial markets. *Journal of Business and Industrial Marketing*, 13(1):8–21.

Swanson Sr, D. R. (2011). Systems, methods and apparatus for self directed individual customer segmentation and customer rewards. US Patent App. 12/567,029.

Syam, N. and Sharma, A. (2018). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. *Industrial Marketing Management*, 69:135 – 146.

Teichert, T., Shehu, E., and von Wartburg, I. (2008). Customer segmentation revisited: The case of the airline industry. *Transportation Research Part A: Policy and Practice*, 42(1):227–242.

Thomas, R. J. (2016). Multistage market segmentation: an exploration of B2B segment alignment. *Journal of Business and Industrial Marketing*, 31(7):821–834.

Walsh, G., Hassan, L. M., Shiu, E., Andrews, J. C., and Hastings, G. (2010). Segmentation in social marketing: Insights from the european union's multi-country, antismoking campaign. *European Journal of Marketing*, 44(7/8):1140–1164.

Westad, F., Hersletha, M., Lea, P., and Martens, H. (2003). Variable selection in pca in sensory descriptive and consumer data. *Food Quality and Preference*, 14(5-6):463–472.

Wind, Y. (1978). Issues and advances in segmentation research. *Journal of marketing research*, 15(3):317–337.

Wu, R.-S. and Chou, P.-H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, 10(3):331–341.

Yun, Y.-H., Li, H.-D., Deng, B.-C., and Cao, D.-S. (2019). An overview of variable selection methods in multivariate analysis of near-infrared spectra. *TrAC Trends in Analytical Chemistry*.

Zadeh, L. A. (1983). A computational approach to fuzzy quantifiers in natural languages. In *Computational linguistics*, pages 149–184. Elsevier.