Toward Interpretable Personas for Banking Customers

by

Md Monir Hossain

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Software Engineering & Intelligent Systems

Department of Electrical & Computer Engineering

University of Alberta

Abstract

The financial landscape is in a state of major disruption through digitalization. The steady adoption of Artificial Intelligence (AI) has brought about a myriad of opportunities for banks and financial institutions to drive efficiency and innovation. These institutions are essentially customer-centric and therefore understanding their customer base is one of the major fields of interest from their perspective. Customer segmentation helps in this by breaking down customers into different groups based on different approaches. In most of the cases, traditional naive approaches like demographic features or specifically calculated financial values are used for this segmentation. The pitfalls of these approaches are the disregard for rich customer data these institutions collect, the introduction of bias, and missing out on the capture of micro-segments. This thesis presents a novel big data analytics framework to create interpretable personas for retail and business banking customers. These data-driven personas are essential to better tailor financial products and improve customer retention.

In this thesis, we start with a comprehensive overview of big data analytics frameworks in finance, time series anomaly detection, customer segmentation, time series clustering, association rule mining, and distributed frameworks in big data analytics and Machine Learning (ML). Then we present the methodology that includes describing the retail and business customer dataset that we use in our experiments. The proposed framework is comprised of several components including pre-processing, anomaly detection, clustering of transaction

time series, and mining association rules that map contextual data to cluster identifiers. We use anomaly detection for improving later stage clustering and find interesting properties for financial time series. We use different raw-data-based clustering techniques and compare them to find out the best methods based on internal evaluation metrics and cluster stability. We then use association rule analysis combining the contextual data with obtained clusters. Thus leveraging rich transaction and contextual data available from nearly 60,000 retail and 90,000 business customers of the financial institution, we empirically evaluate this framework and describe how the identified association rules can be used to explain and refine existing customer classes, and identify new customer classes and various data quality issues. We also analyze the performance of the proposed framework and explain its dynamic nature. We show that it can easily scale to millions of banking customers for both vertical and horizontal scalability.

Preface

The research for this thesis has been led by Dr. Omid Ardakanian in Sustainable Computing Lab and Dr. Hamzeh Khazaei in Performant and Available Computing Systems (PACS) Lab. ATB Financial has collaborated in this research.

This thesis is extended from the workshop paper "Large-scale Data-driven Segmentation of Banking Customers" accepted in the 4th International Workshop on Big Data for Financial News and Data (IEEE Big Data 2020). I was the lead author of the paper. In this thesis my original work includes Chapter 1, Chapter 2, Chapter 3, Chapter 4.1, Chapter 4.2, Chapter 5, and Chapter 6. The ATB Financial team helped in conceptualization and Chapter 4.3 to make sense of the rules. Dr. Omid Ardakanian and Dr. Hamzeh Khazaei helped in conceptualization, editing the manuscript, and provided supervision.

Acknowledgements

I would like to express my gratitude to my supervisors Dr. Omid Ardakanian and Dr. Hamzeh Khazaei for their wonderful guidance and continuous support during the thesis. Apart from being wonderful academic mentors, they were also kind human beings which I hope to be. I would also like to thank ATB Financial for their collaboration throughout the process. The support from Compute Canada through cloud resources that were used to run part of the experiments is also much appreciated.

Having the opportunity to be part of both Sustainable Computing Lab, and Performant and Available Computing Systems (PACS) Lab, I learned immensely from my lab mates. Finally, I would like to thank my family for their encouragement and support throughout my research.

Contents

List of Tables

List of Figures

Acronyms

AI Artificial Intelligence

ARIMA Auto-Regressive Integrated Moving Average

ARMA Auto-Regressive Moving Average

BMU Best Matching Unit

COFUST COpula-based FUzzy clustering algorithm for Spatial Time series

CPU Central Processing Unit

CRM Customer Relationship Management

DBPAM Density-Based Partition Around Medoids

DBSCAN Density-Based Spatial Clustering of Applications

DTW Dynamic Time Warping

ED Euclidean Distance

EDR Edit Distance on Real signals

EPACC Electrical Pattern Ant Colony Clustering

ERP Edit distance with Real Penalty

FFT Fast Fourier Transform

FinTech Financial Technology

FSA Forward Sortation Area

GA Genetic Algorithm

GCP Google Cloud Platform

HBMO Honey Bee Mating Optimization

ICA Independent Component Analysis

ISODATA Iterative Self-Organizing Data-Analysis Technique

JI Jaccard Index

LCSS Longest Common Subsequence

LTV Life Time Value

ML Machine Learning

PAM Partitioning Around Medoids

RAM Random-Access Memory

RFM Recency, Frequency and Monetary value

RFMLC Recency, Frequency, Monetary value, Lifetime, Credit scoring

SAX Symbolic Aggregate approXimation

SBD Shape-Based Distance

SIC Standard Industrial Classification

SOM Self-Organizing Map

SQL Structured Query Language

SVC Support Vector Clustering

VM Virtual Machine

WIT Weighted Items Transaction

Chapter 1

Introduction

1.1 AI in Finance: The Disruptive Technology

In the last decade there has been tremendous progress in the field of Artificial Intelligence. Applications of AI are also picking up the pace on the industry front. It is a game changer for industries like healthcare, banking, transportation, logistics, defense, etc. Whether it is for improving existing process, making sense out of vast amounts of data, mimicking human action or predictive capabilities, the full potential of AI from the application perspective is surely yet to be realized.

Financial industry is one of the most lucrative grounds for Artificial Intelligence applications and has the potential to change the landscape. Financial Technology (FinTech) is expected to reach a market value of $309 Billion by 2022 while the broader financial services market is expected to reach $26.5 Trillion by then [39]. The majority of financial institutions are aware of this AI potential and have started integrating or planning to integrate AI into their workflow. This is resulting in the adoption of organization-wide AI strategies. Some of the most widespread applications right now are fraud detection [12], [32], [90], cash flow prediction [81], [106], churn prediction [8], [99], [109], front office chat bots for customer interaction [5] and personalized recommendations [40]. According to the Boston Consulting Group (BCG), through personalized interactions banks can increase their revenue growth by 300 million dollars for every 100 billion dollars of assets they have [15]. Studies have found that banks can potentially save $447 billion by 2023 from AI applications [18].

A big opportunity in this aspect is to understand the behaviour of the customers that in turn can help these institutions improve existing processes, offer better suited products and drive revenue through new strategies. Banks and financial institutions have a vast amount of data regarding their users but it has been found that traditional banks and financial institutions are not making use of this rich data available to them [39]. This data can be divided into four types [29]:

1. Zero-party data: It is the data that the consumers intentionally share with the financial institutions, explicitly revealing their preferences, interests and intent. This data is generated directly by consumers through surveys, polls, quizzes, contests, questionnaires and preference centers. For example, Business Development Bank of Canada (BDC) uses one-question survey in the website to show relevant content. TD Bank has a marketing preference center from where customers can have granular control of marketing offers or survey requests from the bank.

2. First-party data: It is the data from the customers that is collected, owned, and managed by the institution itself. It can be account related data (transaction, deposit, withdrawal, loan, etc), website, app, Customer Relationship Management (CRM), social media data, etc. It constitutes the major portion of customer data that financial institutions have.

3. Second-party data: It is data that is collected by partnering with another company, thereby obtaining access to the other company's first-party data. ICICI Bank's partnership with Ferrari to offer a co-branded credit card is an example of second-party data where the bank gets information regarding a lucrative segment. Another example is Google Ad data obtained by the institutions.

4. Third-party data: It is the data purchased from companies that collect and format the data from a variety of sources without any direct relationship with the customers. The data sources can be surveys, social

media, websites, etc.

The lack of data usage is specially true from the context of customer segmentation where naive approaches are still employed. We aim to fill this gap by proposing a framework for interpretable dynamic customer segmentation.

1.2 Traditional Approaches to Customer Segmentation

Customer segmentation has long been an essential tool in the retail financial services industry [49]. Traditionally, the customer base is segmented using demographic information such as age, gender, and professions of individuals, alongside analysis of their savings and spending patterns. These segments are then used to tailor financial products to the specific needs of customers, improve customer retention, and create targeted marketing campaigns [64].

The two biggest pitfalls of traditional methods, and of humans in general, are the inability to analyze big data and the prevalence of narrative-driven biases when interpreting results. Modern time series clustering methods provide an opportunity to simultaneously take a granular and wide look at customer spending over time to determine how customers should be segmented [28], [80]. This, in turn, allows a financial institution to compare traditional, human-centred methods to clustering methods for validating the usability of both. For example, a persona used by the financial institution is a "traditional", or an older (60+), typically married individual with large savings (>$200,000) and moderate income (∼$100,000). While we know people like this exist, does their demographic information capture how they transact? This problem is exacerbated when looking at business customers. The demographic features available for businesses are even more limited than those available for individuals, adding complexity for traditional methods, making a data driven approach critical for business customers in particular.

A financial institution can answer the following questions: does contextual information segment business customers in the same way that their spending patterns would suggest? What are the relationships that are captured in

one approach rather than the other? Is there explainability for the differences between traditional and modern approaches? If the answers to these questions are compelling, it gives the financial institution a new tool for segmenting its customers and therefore, a new host of products and services built around these segments.

1.2.1 Narrative-Driven Biases

One issue that is caused by segmenting customers based on pre-existing notions rather than data is that it introduces bias in the process. This is due to fact that banks have some traditional notions of what makes a group of customers separate from another, be it age, education or net annual income. But these features alone are not enough for establishing a nuanced view of customer behaviour. Therefore, an inherent bias is often introduced during the process. Although, the obtained segments still have their usefulness, it especially fails to capture micro-segments of customers.

1.2.2 Evolving Customer Segments

An important property of banking customers is that their needs change over time and so do their earning and spending patterns. The rise of segmentation itself is due to the fact that one size does not fit all customers. This can also be true for the same customer at different time-snaps. This change can be a natural approach or brought about by significant personal or global events. A timely example of such a global event is current COVID outbreak. EY Future Consumer Index on behavior and sentiment has found that customer segments are already evolving due to the pandemic based on their study [37]. This requires the industry to update and adapt their segmentation strategy. Also, based on the adoption of technologies, for example mobile banking, different segments of customers can arise that behave differently. This difference in behaviour can be influenced by their extent of adoption or the time at which they adopt (early or late adopters). In order to capture and understand this changing behaviour with time we need to consider utilizing the rich time series data of these customers. The traditional segmentation overlooks this specificity.

1.2.3 Performance Implications

One of the challenges of using vast amount of data available to these institutions comes from the performance restriction. As the customer number and the time span of the collected data increase, the solution also needs to scale. There is also a need to analyze how much of the data is actually required to make reliable decision for the intended application. The choice of an appropriate solution also depends on the the scale of data and available computing resources. This leads to important trade-offs between cost, performance, ease of implementation and maintainability.

1.3 Automated and Dynamic Customer Segmentation

The financial services industry is currently in a state of major disruption due to advances in AI. Looking at a longer time frame, it is unclear where the line will be drawn between tasks that require a human versus those that do not. In a future where product and marketing design are also automated, customer segmentation is still essential as an input for multi-agent modeling and simulation. A financial institution must navigate the difficult task of segmenting customers in a way that benefits a larger machine learning model even if the results of the segmentation are not clear to us. To do this, it is essential that the expert opinion acts as the prior knowledge that an algorithm can build upon as the entire industry of financial services becomes more automated.

In this thesis, we start down the path of automated dynamic customer segmentation utilizing a large amount of anonymized transaction and contextual data. We examine several methods for clustering the customer base of a financial institution, and compare them using internal validation measures as well as external validation through expert opinion. We use hierarchical clustering to find anomalies and examine how they change over time. We examine the stability of the clusters over various time frames to determine the frequency with which we need to update clusters. We integrate time series clustering with association rule mining to aid in the interpretation of the clusters. Fi-

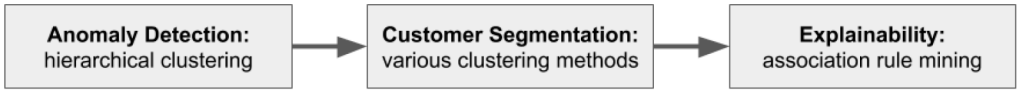

Figure 1.1: Overview of the proposed framework.

nally, we study the vertical and horizontal scalability of the system. Vertical scalability refers to adding more resources to a single computing node while horizontal scalability refers to adding more nodes to the system.

Our methodology involves anomaly detection and clustering of transaction time series, and mining of association rules that map contextual data to cluster identifiers and vice versa. These modules are shown in Figure 1.1 and explained in Section 3.2.2, 3.2.3 and 3.2.4. This three-step approach establishes a framework for developing and deploying interpretable segmentation models across a variety of datasets and timeframes to create actionable insights for the financial institution.

1.4 Summary of Contributions

In this thesis we propose a framework for data-driven customer segmentation. The contribution of this thesis is fourfold:

- We use hierarchical clustering to find and analyse anomalous customers over various time frames. We use raw-data-based time series clustering techniques on a large-scale real banking transaction data from approximately 60,000 retail and 90,000 business customers. We obtain representative patterns for those clusters.

- We investigate cluster stability through different time-snaps and different time horizons.

- We extract useful and interesting rules by combining the metadata (demographic and financial) and the cluster identifier of each customer. In doing so we propose a systematic approach for specifying thresholds for rule mining. We show that these results are helpful for cluster explainability and required for several practical applications in the financial

services industry.

- We evaluate the performance of the automated framework developed for customer segmentation and show that it is vertically scalable. We found that for 1 million customers the pipeline takes 137.62 minutes in Google Cloud Platform (GCP) environment which is viable for the banking institution we partnered with. In addition, we propose a distributed implementation of the framework using synthetic time series data generated by an autoencoder. We therefore confirm vertical scalability through cloud implementation.

1.5 Outline of the Thesis

Chapter 2 discusses the related works on big data frameworks in finance, time series anomaly detection, customer segmentation, time series clustering and association rule mining. Chapter 3 presents the methodology used in the thesis. At first the financial dataset used in this study is described. Then the framework for dynamic customer segmentation is introduced. The preliminary data pre-processing section is described. Next the process of anomaly identification is described which is followed by a comparative analysis of different clustering methods and distance measures used for segmentation. After that a systematic approach for rule mining is proposed.

Chapter 4 presents the experimental result found through the analysis. The stable anomalies found from the time series are presented. Then the best performing clustering techniques are selected based on internal validation and stability consideration. The obtained cluster shapes are also presented. Then the usefulness of the extracted rules are discussed from two different perspectives. One is alignment with existing personas and the other one is the discovery of new personas. Chapter 5 discusses the performance of the proposed framework and shows that it is scalable to millions of customers. Performance for both horizontal and vertical scalability are analyzed along with proposing a distributed implementation of the framework. The context of dynamic nature is then discussed. Chapter 6 concludes the thesis by discussing

the summary of contributions, limitations of the presented work and potential opportunities for future research in this area.

Chapter 2

Literature Review

2.1 Big Data Frameworks in Finance

In the era of big data, the banking industry has gone through radical changes by adopting digitalization. Recent advances in big data mining from the context of the banking industry have been discussed in [50]. It finds that security enhancement, fraud detection, risk management, investment banking, and CRM including customer segmentation are among the top researched topics. There still seems to be a lack of both generalized and application-appropriate frameworks with a focus on real-world application and with consideration of the constraints that banking institutions face. One of the constraints is the confidentiality requirements of banking data that hinders reproducible research in this field. iCARE proposes a non-generalized big data-based customer analytics framework using various data sources [97]. An RFM-based framework for segmentation is proposed in [89] on a small scale of customers. NetDP presents a framework for solving default prediction problem [74]. There is a lack of generalized holistic framework geared towards interpretable segmentation capable of handling time series data.

2.2 Time series Anomaly Detection

Anomaly or outlier detection refers to the deviation from the normal or usual pattern in data. It has a myriad of applications in a wide range of fields including banking and financial organizations [105]. Time series anomaly detection

is a special class of anomaly detection. There have been a number of surveys on the topic of time series anomaly detection and it can be further subdivided as shown in Fig 2.1 [47]. In this study we are specifically interested in whole time series anomaly detection which finds out all anomalous time series among possibly many time series. The goal of this anomaly detection is to obtain better clusters in the clustering stage.

2.3 Customer Segmentation

Different industries, such as financial services [53], have resorted to customer segmentation to better serve their customers and improve customer retention. In most cases, such an analysis is limited to clustering customers based on demographic data or features hand-crafted on a small dataset; this can be seen in Table 2.1 which summarizes applications of clustering to customer segmentation in different industries. Recency, Frequency and Monetary value (RFM) [2], [6], [20], [53], [65], [83], [102] and Life Time Value (LTV) [20], [83] are particularly popular among the hand-crafted features that are used. RFM [55] helps identify more profitable customers. Recency means how recently a customer has transacted, frequency means the number of transactions in a specific period of time and monetary value means the amount of transaction in a specific period of time. On the other hand, On the other hand, LTV estimates average revenue that a customer will generate for the institution. This approach is limiting as it does not utilize rich time series data available from customers. Furthermore, extracting and engineering features manually requires a conscious knowledge of what we are looking for, is prone to various biases, and could reflect stereotypes. Using raw data instead can reveal unexpected and interesting patterns. Also from Table 2.1 we can see that most of the studies are on a rather small number of customers.

Table 2.1: Customer Segmentation in the literature.

Application	Feature	Clustering Method	Customer Size	Paper

Telecom	Current value, potential value and customer loyalty	Decision Tree	2,000	[60]
Automobile	RFM and LTV	Genetic Algorithm (GA)	4,659	[20]
Automobile	RFM	k-means	326	[79]
Fright Transport	Customer survey data	Self-Organizing Map (SOM) and genetic k-means	600	[65]
Airline	Members' travel related attributes like flight count, distance travelled	Density-Based Spatial Clustering of Applications (DBSCAN)	62,988	[54]
Manufacturing	Attributes for Sales and Marketing Department	Genetic k-means	200	[52]
Online Shopping	Customer survey data	SOM	779	[100]
Retail Shopping	Sales log	k-means	10,000	[70]
Tourism	Tourist demographic and stay related information like age, income, length of stay, etc.	SOM & k-means	72,413	[113]
Banking	RFM	k-means & hierarchical	1,904	[2]
Banking	Age, income, deposit, credit, profit/loss	DBSCAN, k-means & two-phase clustering	300	[114]
Banking	RFM, demographic and LTV data	k-means	491	[83]
Banking	RFM & behaviorial scoring	SOM	158,126	[53]
Banking	Date, time, status of transaction, type of transaction, and RFM score	k-means & SOM	2,096	[102]

Banking	customer attributes like demographic and product information	k-means	3,698	[112]
Banking	relative weights of Recency, Frequency, Monetary value, Lifetime, Credit scoring (RFMLC) variables	k-means	4,459	[6]
Banking	customer survey data	k-means & hierarchical	3,480	[92]

2.4 Time Series Clustering

Time series clustering is an important unsupervised learning technique which has been widely utilized to uncover hidden patterns in a given dataset. Traditional clustering algorithms with different distance (or similarity) metrics are commonly used for time series clustering [72].

2.4.1 Clustering Algorithms

In general, time series clustering techniques can be divided into raw-data-based and feature-based approaches [34]. In the raw-data-based approach, time series data are directly used as input to an appropriate clustering algorithm. For banking customers, this can be time series of transaction amount or frequency. The feature-based approach is however based on application-specific features that are extracted from time series and then used as input to the clustering algorithm. These features can also be parameters extracted from a fitted model. For banking customers, this can be their calculated book value or parameters of an Auto-Regressive Integrated Moving Average (ARIMA) model which is fit to the transaction amount time series.

The feature-based approach is robust to noise and can reduce the dimension of data, but it assumes the availability of high-quality features. Additionally, the time complexity of these techniques is usually higher than raw-data-based techniques due to the complexity of model fitting and feature extraction. There

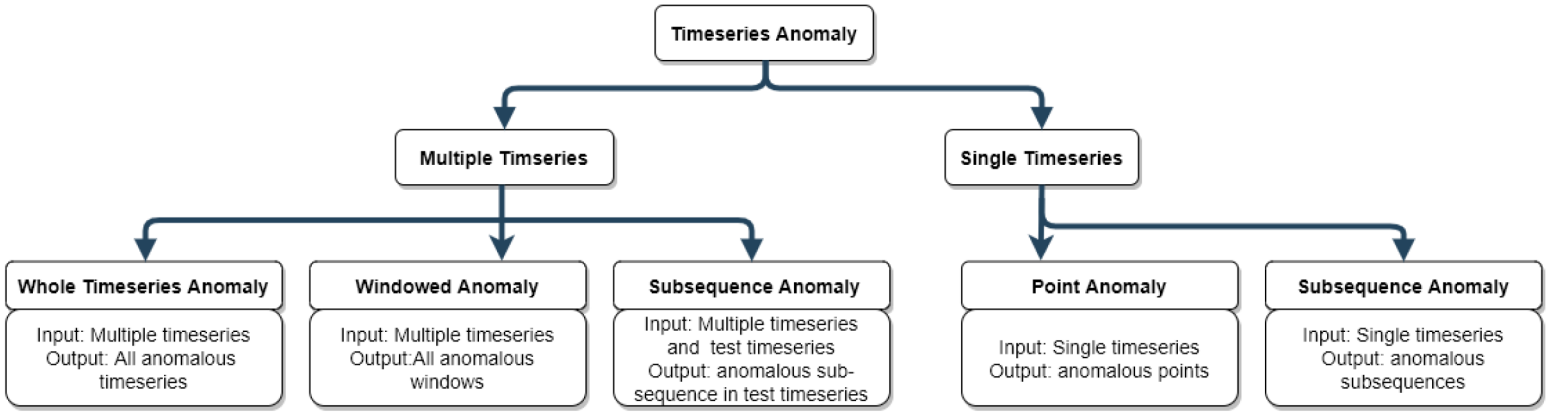

Figure 2.1: Classification of time series anomaly detection.

are two types of feature-based methods: those that perform feature extraction and clustering separately and those that jointly optimize the feature learning and clustering [76]. We summarize feature-based clustering in Table 2.2.

We see a prevalence of studies with electrical energy time series in terms of application area. A number of the studies extract features using statistical and domain-specific calculations, or parameters extracted from fitted models. For example, derived statistical features like mean, skewness, kurtosis, etc from hourly load data with a one-week window are calculated in [94]. This is then used for clustering with the application being the creation of more accurate load curves for small customers. Reference [82] uses pattern vectors formed using load model and weather parameters for clustering. The result is used for customer classification to use in distribution network calculation. Reference [46] uses weights/mixing matrix from Independent Component Analysis (ICA) of Chinese stock return data. The goal of clustering here is to better understand the stock trends. Reference [38] proposes an electricity consumer characterization framework to help the retail and distribution companies in extracting useful information from electricity consumption data. It involves a data reduction stage using previous knowledge about the way the loading conditions (e.g., the season of the year and the type of weekday) affect electricity consumption and dimensionality reduction using SOM. Reference [86] proposes a method for fuzzy classification of load demand profiles to be used in tariff development and end-user cost determination. This study first uses the cosine amplitude method to produce a fuzzy relation matrix and then uses the max-min composition method for generating a fuzzy equivalence matrix. This matrix is then used for clustering. Reference [66] studies stock market industrial sectors interconnection dynamics using clustering based on Hurst estimated exponents. Reference [93] uses normalized load profile for intrinsic cluster analysis and combines it with Standard Industrial Classification (SIC) code, tariff code, and load factor for extrinsic cluster analysis. Reference [115] uses low dimensional stock data for clustering to obtain a holistic insight into the development of assets, market sectors, countries, and the financial market as a whole. References [25], [26] use time series data, direct and indirect shape

features extracted from time series data, and harmonics-based coefficients as features to understand behaviour of electricity customers. Reference [84] uses load profile alignment coefficients calculated from the average load at different periods of time and daily average load for clustering to improve load modeling accuracy. Reference [56] uses the coefficient of variation in daily loads, the sum of coefficients of variation in hourly loads, the sum of standard deviations in hourly load shares, and the sum of coefficients of variation in hourly load shares as features. The goal of clustering customers here is to better predict how they would respond to a demand response program. Reference [107] uses the weekly closing price of stocks for clustering to properly identify industry category. Reference [31] uses total usage during the evening period, the variability of time of maximum usage during the evening, and variability of time of minimum usage during the evening as features to group households together. Reference [87] extracts frequency domain coefficients using Fast Fourier Transform (FFT) while References [66], [108] use wavelet coefficients as features. Reference [85] uses Symbolic Aggregate approXimation (SAX) which is a symbolic representation for time series that reduces data dimension through a synthetic set of symbols. Reference [58] uses partial correlation of stock time series to group firms and study how they related to others over different periods of times. Reference [13] estimates the Auto-Regressive Moving Average (ARMA) model parameters from the data and uses them for clustering. As we can see a wide variety of extracted features are used based on the applications.

Table 2.2: Feature-based time series clustering the literature.

Application	**Feature**	**Clustering Method**	**Cust. Size**	**Ref.**
Electrical energy	Mean, standard deviation, skewness, kurtosis, chaos, energy and periodicity	k-means	1,035	[94]

Financial	Features extracted by ICA from stock	Modified k-means	30	[46]
Electrical energy	Load model & weather parameters	Iterative Self-Organizing Data-Analysis Technique (ISODATA)	660	[82]
Electrical energy	Dimension reduced load data by SOM	k-means	165	[38]
Electrical energy	Fuzzy relation matrix from load	Lambda-Cuts for fuzzy relation method	365	[86]
Financial	Hurst exponent estimates of stocks	Hierarchical clustering	8	[66]
Electrical energy	Normalized load profile, day of the week and month, load factor, tariff code and SIC	Hybrid decision tree clustering	500	[93]
Financial	Low dimensional stock data	k-means	46,237	[115]
Electrical energy	Time domain data, shape factors & harmonics based coefficient	Modified 'follow the leader' procedure, Two variants of hierarchical clustering, SOM, k-means, fuzzy k-means	234	[25]
Electrical energy	Shape indicator of load profile	Modified follow the leader	471	[26]
Electrical energy	Load profile alignment coefficients	Statistical clustering	26	[84]
Electrical energy	Four variability indices from load profile	Hierarchical k-means	802	[56]
Financial	Weekly closing price of stock	Hierarchical agglomerative clustering	91	[107]
Electrical energy	Total usage measure, flexibility measure for time of maximum and minimum usage	k-means	180	[31]

Electrical energy	Frequency-domain parameters obtained from load profile using FFT	k-means++, SOM	6,570	[87]
Electrical energy	SAX representation of 15 minute load profile	Agglomerative hierarchical clustering	234	[85]
Financial	Wavelet Transform based co-efficients of stocks	Hierarchical clustering	9	[66]
Financial	Correlation coefficients of stocks	Agglomerative clustering	732	[58]
Electrical energy	Time domain representation of load profile using discrete wavelet transform	k-means	18,000	[108]
Economic	Colupas	COpula-based FUzzy clustering algorithm for Spatial Time series (COFUST)	116	[35]
Health	ARMA Parameters	k-means, k-medoids	70	[13]
Loan	Relative savings amount time series	Hidden Markov Model-Based Clustering	50,000	[62]

To the best of our knowledge, time series clustering has not been used to group banking customers using their raw transaction time series data. Although there is a study that used time series data of savings and loan data for feature-based clustering [62]; this work assumes the existence of an underlying model and is computationally expensive. As we are specifically interested in transaction patterns of customers along with segmentation, we adopt a raw-data-based approach in our work. A summary of raw-data-based approaches is provided in Table 2.3.

Table 2.3: Raw-data-based time series clustering the literature.

Application	**Feature**	**Clustering Method**	**Time Series Size**	**Ref.**
Electrical energy	Hourly energy consumption	Iterative refinement clustering	234	[14]
Electrical energy	15-minutes energy consumption	SOM followed by k-means	165	[69]
Social	Social media usage	k-medoids	17,231	[21]
Electrical energy	15-minutes energy consumption	k-means & hierarchical	235	[59]
Electrical energy	Hourly energy consumption	Fuzzy c-means & hierarchical	288	[44]
Electrical energy	15-minutes energy consumption	k-means, Kohonen adaptive vector quantization, fuzzy k-means, and hierarchical clustering	94	[98]
Electrical energy	Customer base load	k-means	1,182	[110]
Electrical energy	Hourly energy consumption	k-shape	29,280	[111]
Health	Average count of mosquito eggs per week	k-shape, Partitioning Around Medoids (PAM), k-DBA (DTW barycenter averaging)	300	[9]
Electrical energy	Hourly energy consumption	Weighted fuzzy average (WFA) k-means	316	[78]
Electrical energy	Hourly energy consumption	Fuzzy c-means & hierarchical	283	[44]
Electrical energy	15-minutes consumption	Support Vector Clustering (SVC)	234	[22]
Electrical energy	Hourly load profile	SOM, modified follow the leader	31	[24]
Transportation	Passenger service rate	Hierarchical clustering	64	[38]

Electrical energy	15-minutes energy consumption	Fuzzy c-mean	310	[42]
Electrical energy	15-minutes energy consumption	Modified Electrical Pattern Ant Colony Clustering (EPACC)	234	[23]
Transportation	Hourly occupancy of parking station	Density-Based Partition Around Medoids (DB-PAM)	27	[71]
Electrical energy	Hourly energy consumption	Honey Bee Mating Optimization (HBMO)	N/A	[41]
Weather	Weekly SO_2 concentration	Fuzzy k-medoids	65	[33]
Electrical energy	15-minutes energy consumption	Fuzzy c-means	513	[43]
Electrical energy	Half-hourly energy consumption	k-means	100	[101]
Electrical energy	15-minutes energy consumption	Fuzzy-neural constructive and merging hierarchical algorithms	365	[67]

2.4.2 Distance Measures

Clustering methods rely on a measure of (dis)similarity to calculate the distance between data points. The most popular distance metric is the Euclidean Distance (ED) which computes the 2-norm of the difference between two data points. To use this distance metric all the time series need to be of equal length, otherwise, they need to be trimmed or concatenated. Often time, transaction data are not of equal length as customers can open an account with a financial institution at different points in time. For time series of varying lengths another well-known metric, called Dynamic Time Warping (DTW) [96], can be helpful. DTW calculates the optimal match of two-time series instances with certain constraints and offers invariance to certain distortions. It has been used to find patterns in time series [17] and also as a distance metric in time

series clustering [45]. Shape-Based Distance (SBD) is another distance metric used in k-shape clustering [88].

A previous study [104] found that for large datasets, ED yields an accuracy closer to that of elastic measures like DTW, Longest Common Subsequence (LCSS), Edit Distance on Real signals (EDR), and Edit distance with Real Penalty (ERP). Among the elastic measures, DTW and constrained DTW (cDTW) perform well comparatively. In this study, we use ED, DTW and SBD in different clustering algorithms.

2.5 Association Rule Mining

Association rule mining is a popular pattern discovery method originally used in the database literature [3]. Different algorithms have been proposed for association rule mining, such as Apriori [4] and FP-growth [48]. Apriori is the most popular implementation among these algorithms [51] and has several variations. The financial services industry has used demographic and transactional features to obtain interesting rules about their customers using fuzzy [11], apriori [53] and Weighted Items Transaction (WIT) tree [10] based association rule mining techniques.

Class association rules are a special type of association rules where the antecedent of the rule is a set of features and its consequent is the corresponding class label [75]. In this work we integrate time series clustering with association rule mining to extract interesting and useful rules where the antecedent is metadata elements and the consequent is the cluster identifier and vice versa. A summary of the application of association rule mining in Banking is provided in Table 2.4.

Table 2.4: Association rule mining application in the literature.

Application	**Feature**	**Rule Mining Method**	**Cust. Size**	**Ref.**
Banking	Demographic, account and transaction information	Fuzzy association rule mining	320,000	[11]

Banking	Features in transaction database	Apriori association mining	158,126	[53]
Web/ banking	17 phishing characteristic indicators extracted from phishing url	Apriori and predictive apriori association mining	1,400	[1]
Banking	Demographic Feature	Association rule mining using WIT tree	181	[10]

2.6 Distributed Frameworks in Big Data Analytics and ML

The need for distributed frameworks in data analytics and machine learning stems from the need to process huge of amount of data that that can not be processed by horizontally scaled resources or decentralized data that can not be brought together due to different constraints. Several open-source frameworks are available for this purpose which have their advantages and disadvantages [68]. These implementations can be divided into the frameworks and their processing engines. Some popular processing engines are Hadoop, Apache Spark, Apache Storm, Flink, H_2O and Samza and popular ML frameworks include MLib, Flink-ML, H_2O ML Library, Mahout, Oryx, SAMOA and GraphLab. In this study we use Spark and MLib which has the capability to handle both batch and streaming (through micro-batching) data.

Chapter 3

Methodology

In this chapter we describe our methodology for finding, explaining, and evaluating association rules after anomaly detection and clustering as depicted in Figure 1.1. This part of the implementation is done within the GCP. We, therefore, use different components of GCP like BigQuery and AI Platform. The data ingestion is done from BigQuery while the anomaly detection, clustering, and association rule mining are performed in the AI Platform.

3.1 Exploratory Description of Retail and Business Customers

Our dataset is divided into two parts, namely, transaction data and metadata. Combined it contains approximately 428 million data points from the customers for each month.

3.1.1 Time Series Transaction Data

The transaction data contains fully anonymized debit transaction information for approximately 300,000 retail customers and 90,000 business customers of a financial institution for 2 years.

To decide on the clustering algorithm, distance measure, and rule mining thresholds, we take a random sample of 20% of the retail customers which yields a reasonably sized dataset for fast experimentation. We only consider spending transactions (money flowing out of an account) aggregated into number of transactions per day for each customer. The sampled dataset contains

the transactional data for 59,370 retail customers. For business customers, we use the full dataset. We initially segment the retail customers and also use the transaction data of the same set of customers during other time periods for stability analysis. Through this, we identify the two best-performing algorithms which we then use for business customer segmentation. To study the performance and scalability of the proposed framework, we run the same two algorithms on all 300,000 retail customers.

3.1.2 Contextual Metadata

For the second part, we have metadata for the selected retail and business customers. This metadata is linked to the transactions data via an anonymous ID, and all sensitive fields are hidden. Tables 3.1 shows the features that are available in the retail customer metadata. As we can see there are missing values within the data. We do not modify or impute them and let the association rules be discovered in the presence of these missing values.

Table 3.1: Metadata summary for retail customers.

Feature	**Data Type**	**no. Bins**	**pct. Available**
Age	Numerical	4	84.53
Gender	Categorical	3	100.00
Post Tax Income	Numerical	4	64.22
Forward Sortation Area (FSA)	Categorical	2	93.12
Has Joint Partner?	Binary	2	100.00
Product Count	Numerical	4	100.00
Book Value	Numerical	4	100.00

We use the metadata for association rules mining to obtain rules regarding the segments and explain existing personas. We also use this dataset to confirm that the selected random sample is representative of the whole customer base.

To make sure that the sampled retail customers are representative of the whole dataset, we compare the distributions and ratios of categorical values between the full dataset and the sampled population. For continuous variables, we compare the distributions using KS statistics [73] as shown in Table 3.2.

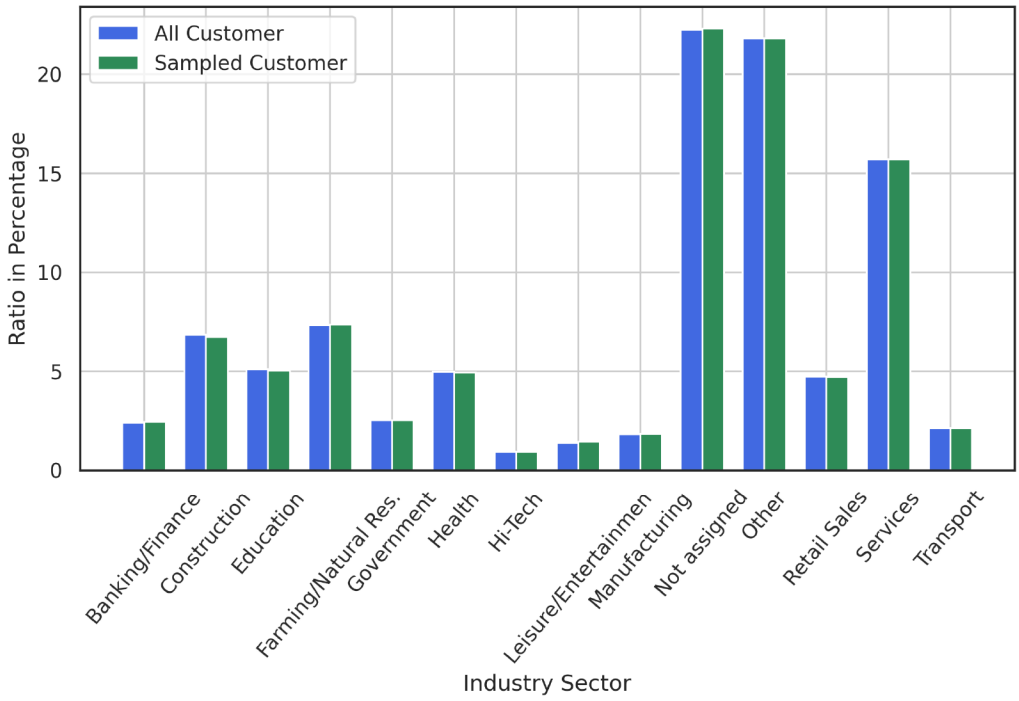

Figure 3.1: Industry sectors that the retail customers belong to.

Since the p values are more than 0.05, the null hypothesis that the distributions are similar cannot be rejected. The fact that both the distributions have a very large number of samples, the result becomes more conclusive.

Table 3.2: KS statistics table for continuous variables.

Feature	**Statistic**	**p value**
Age	0.0034	0.5654
Net Annual Income	0.0039	0.2956
Book Value	0.0025	0.8228

For categorical features, e.g., FSA, Gender, and Has Joint Partner (or not), we checked the ratio of different categories and found them to be similar. Figure 3.1 shows that the ratios of different industry sectors in the full dataset are nearly identical to those in the sampled dataset. Based on these results, we argue that the sampled dataset provides a reasonable representation of the full dataset.

Table 3.3 shows the features that are available for business customers' rules mining. The unique metadata available here is the NAICS code, which defines industry type. Here, we use a two-digit NAICS code, which defines the highest level of industry type.

Table 3.3: Metadata summary for business customers.

Feature	**Data Type**	**no. Bins**	**pct. Available**
NAICS Code	Categorical	25	100.00
FSA	Categorical	3	66.91
Book Value	Numerical	4	100.00
Post Tax Income	Numerical	4	81.28
Product Count	Numerical	3	100.00

3.2 A Framework for Dynamic Customer Segmentation

In our solution, we propose a framework for scalably deploying a trained unsupervised learning model into a production environment at a financial institution. Our solution aligns with the existing architecture being utilized by the financial institution, which primarily uses the Google Cloud Platform, both for data storage (BigQuery) and model training and deployment (AI Platform). The solution can also be translated to other cloud platforms.

The system architecture can be seen in Fig 3.2. In this pipeline, as transactions appear in the financial institution's source platform, they are automatically streamed to the BigQuery Structured Query Language (SQL) environment. Data pre-processing is done in BigQuery. For transactions data, this means aggregation while for metadata this means binning and aggregation. The batch transactions data is then loaded into the AI platform that may also contain a pre-processing module if necessary. Here at the beginning, the anomaly detection module segregates the anomalies and passes the data to the segmentation module for clustering. Then this result is combined with metadata loaded from BigQuery and generates association rules for explainability. The labeled customers are then once again loaded into BigQuery to

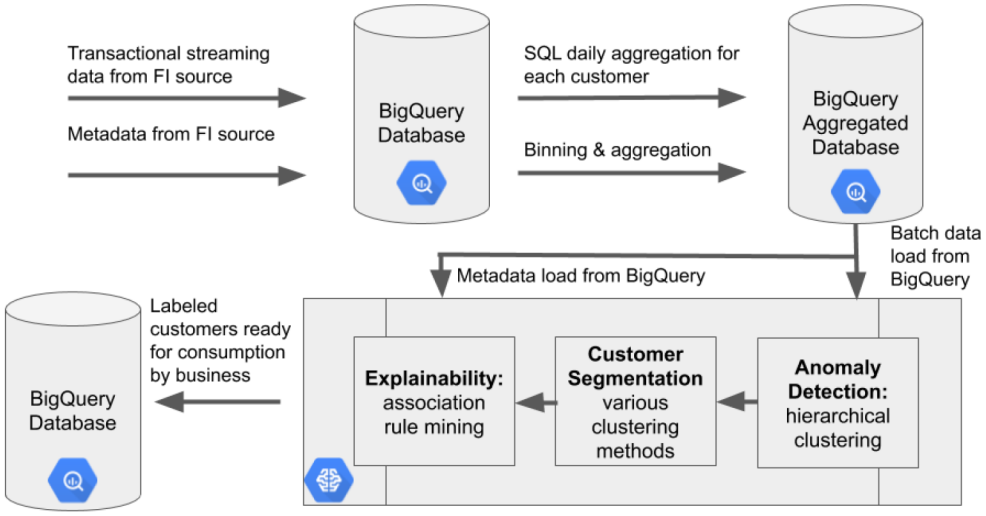

Figure 3.2: Architecture of the implemented big data analytics pipeline.

be consumed by business.

3.2.1 Data Pre-processing

For all transaction datasets, we convert the transaction data into daily transaction count. This allows us to compare clustering methods that adopt different distance measures. For rule extraction, based on the input from the domain experts, we map numerical values into a certain number of ranges using a quartile-based binning strategy, yielding 4 bins. The ranges in these bins make sense from the business perspective as per the domain experts. However, for business customers' product count there are not enough distinct values for 4 quartile-based bins, so we select 3 bins. For the FSA codes, we convert them to rural and urban areas.

3.2.2 Anomaly Identification

We use agglomerative hierarchical clustering for whole time series anomaly detection in both retail and business customers. This is a suitable algorithm for anomaly detection as it creates a dendrogram that can be cut at different heights to control the numbers of anomalies. We found it is able to detect

appropriate anomalies for the intended purpose, i.e., to prevent singletons in later clustering stage. It is however possible to compare it with other anomaly detection algorithms like density based ones. Agglomerative hierarchical clustering is a type of hierarchical clustering [57]. Here each sample represents a cluster of its own at the beginning and then those clusters are gradually combined. The metric used for merging depends on the selected linkage method. The output of the algorithm is a dendrogram that represent how the samples are merged and at which level. By cutting this dendrogram at different levels, different numbers of clusters can be obtained. For our use case, as the anomalous samples are quite distinct compared to normal samples, they appear as singletons after the cut. We experiment with different linkage methods, length, and shift of data. The goal is to select the most suitable linkage method and length of data along with figuring out if customer data which is shifted in time to some extent can be used by the framework.

3.2.3 Comparative Analysis of Clustering Techniques

We use k-means, k-medoids, k-shape, and SOM as our clustering methods with the sampled retail customer dataset. We use Euclidean distance and DTW as distance metrics with k-means and k-medoids algorithms, while for SOM we use only Euclidean distance and for k-shape we use shape-based distance. We run the algorithms for 2 to 20 clusters and calculate the normalized sum of squared distances (i.e., the distortion score). We select the number of clusters to be 5 using the elbow method as shown in Figure 3.3. For business customers, we use k-means with Euclidean distance and SOM as they proved to be the best performing algorithms. In this case, we set the cluster number to 6 based on the knee of the SSE curve.

Euclidean Distance (ED)

It is the distance between two points in Euclidean space. Let $x = [x_1, \ldots, x_n]$ and $y = [y_1, \ldots, y_n]$ be two points in an n-dimensional Euclidean space. The Euclidean distance between these two points is defined as:

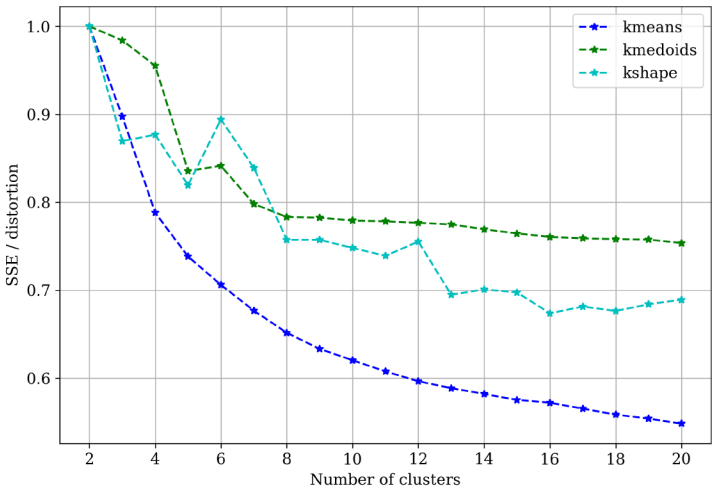

Figure 3.3: Normalized SSE/distortion of clusters.

$$\mathrm{ED}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Dynamic Time Warping (DTW)

It allows comparison between two univariate time series. Let $x = [x_1, x_2, \cdots, x_m]$ and $y = [y_1, y_2, \cdots, y_n]$ be two sequences of length m and n, respectively. An $m \times n$ matrix can now be formed where each element of this matrix corresponds to an alignment between points (x_i, y_j). The goal is to find a warping path $w = [w_1, w_2, \cdots, w_k]$ of length k which minimizes the distance between the two sequences:

$$\mathrm{DTW}(x, y) = \min_{w} \sum_{i=1}^{k} \delta(w_i)$$

where δ is a suitable distance measure between the points (x_i, y_j). For example, this distance measure can be the 1-norm of the difference between these two points $\delta(i, j) = |x_i - y_j|$. The warping path itself is usually subject to a set of constraints, including boundary conditions, monotonicity, continuity, warping window, and slope constraints [17].

Figure 3.4 shows the alignment between two time series x and y based on DTW. Here as opposed to Euclidean distance's point-to-point alignment as

shown by the dotted lines, DTW finds the best alignments between the two time series. The distance matrix between the two time is shown in Figure 3.5. The darker grids are the most aligned points and therefore constructs the optimum warping path.

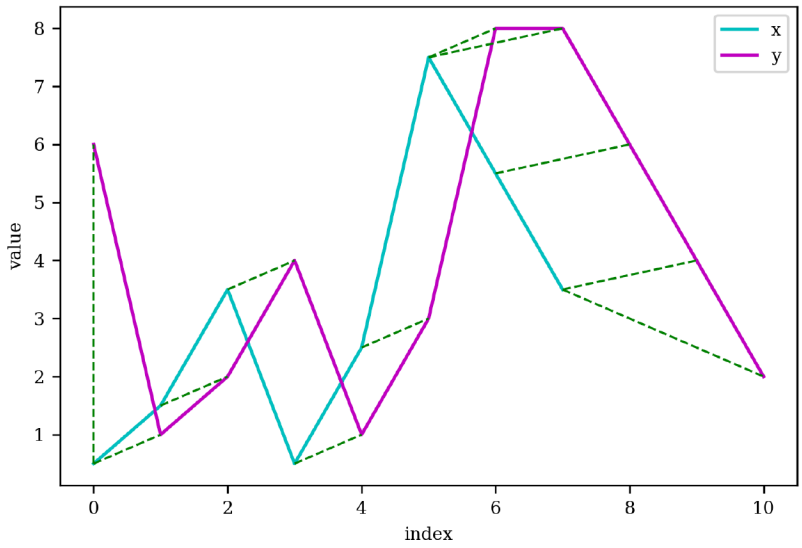

Figure 3.4: Alignment between two time series based on DTW.

k-means

The k-means algorithm is a widely used partitioning-based clustering method [77]. Consider a set of p points in $\mathbb{R}^n$ represented as $X = \{x_1, \cdots, x_p | x_i \in \mathbb{R}^n \quad \forall i\}$ which is to be partitioned into k clusters where the cluster centers are given by $C = \{c_1, \cdots, c_k | c_i \in \mathbb{R}^n \quad \forall i\}$. In effect, the algorithm solves the following optimization problem and is guaranteed to converge in a finite number of iterations:

$$\min \sum_{k=1}^{K} \sum_{i=1}^{n_k} ||x_{i,k} - \mu_k||_2^2$$

Here K is the total number of cluster, n_k is the number of members in the kth cluster, $x_{i,k}$ is the ith member of this cluster, and μ_k is the center of this cluster.

Although ED is the distance measure commonly used with k-means, DTW

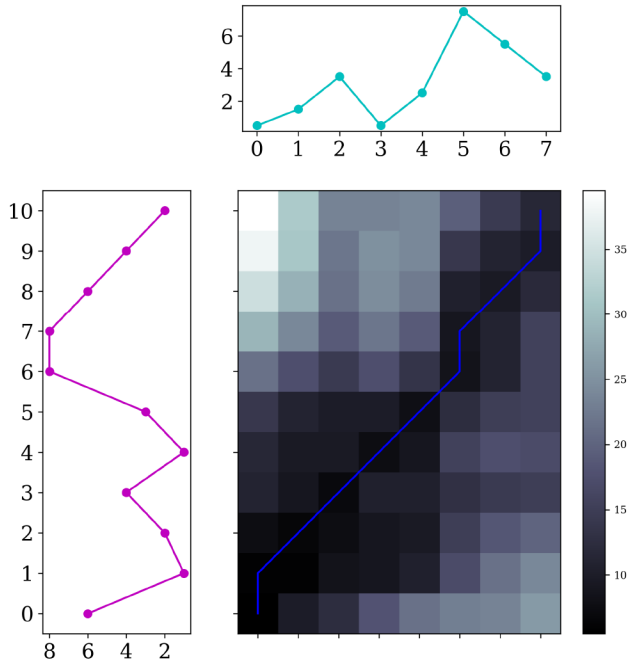

Figure 3.5: A dynamic time warping path example.

can also be used with k-means clustering. In this case, the cluster centers are calculated using DBA [91] and the variation of the algorithm is known as k-DBA.

k-medoids

The k-medoids algorithm is another popular partitioning-based algorithm [36]. It is similar to k-means in the sense that it minimizes the sum of squared errors between the cluster center and cluster members, i.e., they have the same objective function. However, instead of calculating a virtual cluster center from the average of cluster members, it considers a real data point as the cluster center. The k-medoids algorithm can be used with both ED and DTW distance measures.

k-shape

The k-shape algorithm is a partitioning-based clustering method which uses a shape-based distance (SBD) measure and is designed specifically to preserve

the shape of time series [88]. It is invariant to scaling and shifting, and is faster than algorithms that use the DTW distance measure. Scale invariance refers to the case when two time series are of the same shape but have different magnitudes and/or lengths, whereas shift invariance refers to the case where two time series are similar in shape but differ in phase. It uses a normalized version of cross-correlation as the distance measure to decide the centroid of each cluster and then updates the members of each cluster. For the time series x and y of length n, a cross-correlation sequence of length $2n-1$ is calculated as follows:

$$CC_w(x,y) = R_{w-n}(x,y), \quad w \in \{1,\ldots,2n-1\}$$

where

$$R_k(x,y) = \begin{cases} \sum_{l=1}^{n-k} x_{l+k} \cdot y_l, & k \geq 0 \\ R_{-k}(y,x), & k < 0 \end{cases}$$

The distance measure is given by

$$\text{SBD}(x,y) = 1 - \max_w \left(\frac{CC_w(x,y)}{\sqrt{R_0(x,x) \cdot R_0(y,y)}}\right)$$

where w is the position at which the normalized cross correlation is maximized.

Self Organizing Map (SOM) [63]

The Self Organizing Map is a type of artificial neural network that projects n-dimensional input data into a lower-dimensional grid space while preserving the topology in the projection. Each neuron has a specific topological position and a vector of weights, known as a codebook vector, which has the same dimension as input data and is randomly initialized. The codebook vector of neuron i is represented by $m_i = [m_{i1}, m_{i2}, \ldots, m_{in}]$. When a random input x is fed to this network during training, the distance (ED or any other distance measure) between this input and each of the codebook vectors is computed. The neuron with the minimum distance is called the Best Matching Unit (BMU). Thus, if m_c is the BMU then,

$$||x - m_c|| = \min_i \{||x - m_i||\}$$

Subsequently, the weights of the neurons, i.e., the codebook vectors, are updated using this formula:

$$m_i(t+1) = m_i(t) + \alpha(t)h_{ci}(t)[x(t) - m_i(t)]$$

where t is the time index, c is the index of BMU, and $\alpha(t)$ is the learning rate which decreases with time but always remains between 0 and 1. $h_{ci}(t)$ is called the neighbourhood function and is used to describe the neighbourhood area of a neuron. Specifically, it is a function of the grid-distance between the neuron c and neuron i. A popular choice for this function is a Gaussian. This process is repeated several times for each input.

Once the SOM is trained, it assigns an input vector to a cluster by finding the neuron which has the minimum distance from it. Hence, the number of neurons determines the number of clusters.

3.2.4 Systematic Approach Towards Extracting Useful Rules

Consider a dataset D of n records: $D = \{T_1, T_2, T_3 \cdots, T_n\}$ describing n customers. Each record $T_i \in D$ is a set of items (i.e., features), hence it is called an *itemset*. Let I denote the set of all items. We can write $T_i = \{I_{i_1}, \cdots, I_{i_k} | I_{i_j} \in I, \quad \forall j\}$. The rule $A \Rightarrow B$ expresses that if a record contains A then it probably contains B too. The implicit assumption here is that $A, B \subseteq I$ and $A \cap B = \emptyset$.

Association rules must satisfy a minimum support constraint and a minimum confidence constraint at the same time. The support of an itemset is the probability that this itemset appears in D.

$$\text{support}(A) = \frac{|\{T \in D; A \subseteq T\}|}{|D|}$$

Here $|\mathcal{D}|$ denotes the cardinality of a set $\mathcal{D}$. The support of rule $A \Rightarrow B$ is defined as support$(A \cup B)$ which is the probability of co-occurrence of A and B in D. The confidence of this rule is the ratio of the number of records in D that contain both A and B to the number of records that contain A. It can

be interpreted as an estimate of $P(B|A)$.

$$\text{confidence}(A \Rightarrow B) = \frac{\text{support}(A \cup B)}{\text{support}(A)}$$

Note that $\text{support}(A \cup B)$ can be viewed as the probability that a record contains A in addition to B.

Association rule mining consists of two steps: identification of frequent itemsets and generation of rules [27]. The frequent itemset generation is a computationally expensive task. Several algorithms have been proposed to efficiently identify the frequent itemsets and in this work we use the Apriori algorithm [4].

We use the metadata described in Section 3.1.2 for association rule mining. In particular, each record in this table is contextual data about a specific customer. After clustering customers based on their transaction data, we add a column to the metadata table to store the cluster (ID) that each customer is assigned to. We only consider rules that have one of the following forms:

$$\text{Cluster } k \Rightarrow \text{Feature set}$$

$$\text{Feature set} \Rightarrow \text{Cluster } k$$

where each feature set includes one or more metadata properties. We refer to these as Type-1 and Type-2 respectively.

Our goal is to find both Type-1 and Type-2 association rules that show a higher correlation that one would expect by random chance even if the rules apply to only a handful of customers. This motivates us to use a small support threshold of 0.1% to get at least a handful of rules even for the smallest cluster. Confidence thresholds are selected using a systematic approach. We start by determining the proportion of the antecedent that we would expect to contain the consequent by random chance. For Type-1 rules, this is the co-occurrence of the features within the full dataset. The second step is to multiply the found probabilities by an experimental constant that results in the confidence being significantly larger than we would expect by random chance. We set the experimental constant to be 1.5. For Type-2 rules, this is the proportion of the cluster size to the size of the dataset. Hence, the thresholds are: 0.9 for

large ($\geq$ 60%), 0.8 for medium ($\geq$ 30%), 0.6 for small ($\geq$ 15%), and 0.3 for tiny (<15%) clusters.

It is important to note that using this methodology, Type 2 rules are particularly important. While Type 1 rules explain a cluster, Type 2 rules show that a significant portion of the population defined by the features is contained in a single cluster.

Chapter 4

Experimental Results

4.1 Stable Anomalous Customers for Transaction Time Series

We find that different linkage methods give us very similar time series anomalies. With a fixed distance threshold of 100, we obtain similar number of anomalies as seen in Table 4.1. Single linkage ends up with the least anomalies. However, 23 of the anomalies that it identifies are common in other linkage methods too. Table 4.1 also shows run-time for different linkage methods. Single linage is significantly faster compared to the rest with being 73.22% and 71.59% faster than the closest ones. As in different implementations there will be the need to deal with larger scale of data (e.g., retail customers) and the linkage methods identify similar anomalies, we choose single linkage for the next steps of the experiments given it's faster runtime.

Table 4.1: Extracted anomalies and runtime in minute for different linkages

Linkage	**Retail Customer**		**Business Customer**	
	No. Anomaly	***Runtime***	***No. Anomaly***	***Runtime***
ward	30	16.56	8	5.43
single	24	3.06	6	1.23
complete	33	14.26	8	5.17
average	34	15.93	7	5.73
weighted	33	15.31	8	5.52
centroid	31	12.13	7	4.58
median	30	11.46	8	4.33

While running the experiments on different lengths, i.e. 1-month, 3-month, and 6-month worth of data, we find that similar anomalies are obtained. This shows that for financial timesries, whole time series anomalies are rather stable over different lengths. Based on this we opt to use 1-month of data as it requires less computation and memory resources.

We also experimented if different starting points of same length of data result in significantly different anomalies. We used 31 days of data starting from 1 January, 8 January, 15 January, and 1 February, 2019. The numbers of obtained anomalies are shown in Table 4.2 21 of these were similar across the 4 starting points. This shows that within a reasonable time shift as long as the same length of data is available, the customer data can be fed into the anomaly detection module. This is especially useful as new customers join the institution and need to be included in the analysis.

Table 4.2: Detected anomalies across different starting points of time.

Amount of Data	**Nb. of Anomalies**
Jan 1 - Jan 31	24
Jan 8 - Feb 7	23
Jan 15 - Feb 14	24
Feb 1 - March 3	24

Based on these 3 sets of experiments we decide to do anomaly detection using 1 month of data and single linkage. A few of the extracted anomalies can be seen in Fig 4.1.

This improves the clustering step. For example, without this anomaly detection part, we get 5 singletons in the k-means clustering step with 20 clusters.

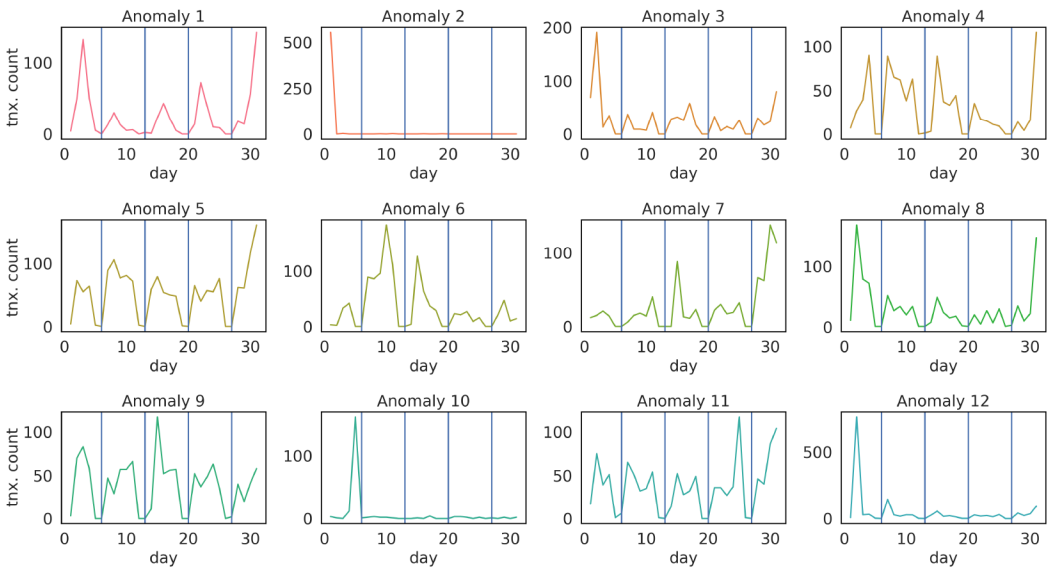

Figure 4.1: Examples of detected time series anomalies.

4.2 Best Performing Clustering Technique

Clustering results can be evaluated using internal, external, and relative metrics. External evaluation is supervised and requires ground truth or expert knowledge. When this expert knowledge is not available internal evaluation metrics can be useful.

4.2.1 Internal Validation

Internal evaluation is unsupervised and uses the inherent properties of data itself for evaluation. We use *Davies-Bouldin* index (DB) [30], *Calinski-Harabasz* index (CH) [19] and *Silhouette Score* (SS) [95] as internal evaluation metrics.

DB index

DB index measures how dense the clusters are and how well-separated they are from one another. It takes into account the intra-cluster diversity (a measure of dispersion) and inter-cluster distances. The intra-cluster diversity is the distance between the members of a cluster and its center. The inter-cluster distance is the distance between the cluster centers. A lower DB index indicates a better clustering.

To calculate DB index, a measure R_ij is calculated as

$$R_{ij} = \frac{S_i + S_j}{M_{ij}}$$

where S_i (S_j) is the average distance between each point of cluster C_i (C_j) and the center of the respective cluster, and M_{ij} is the distance between these two cluster centers. Assuming that there are K clusters, DB index can be obtained as follows:

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{i \neq j} R_{ij}.$$

CH index

CH index can be defined as the ratio of the sum of inter-cluster dispersion and intra-cluster dispersion for all clusters. A higher CH index indicates a better clustering. It can be calculated as,

$$CH = \frac{\text{tr}(B_K)}{\text{tr}(W_K)} \times \frac{N - K}{K - 1}$$

where K is the number of clusters, N is the total number of samples, $\text{tr}(W_K)$ is the trace of the intra-group dispersion matrix, and $\text{tr}(B_K)$ is the trace of the inter-cluster dispersion matrix. These matrices are defined below

$$\text{W}_K = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{i,k} - \mu_k)(x_{i,k} - \mu_k)^\top$$

$$\text{B}_K = \sum_{k=1}^{K} n_k (\mu_k - \mu)(\mu_k - \mu)^\top$$

Here μ is the center of all points in the dataset, μ_k is the center of cluster k, $x_{i,k}$ is the ith point in that cluster, and n_k is the number of points in that cluster.

Silhouette Score

Silhouette Score describes how similar a sample is to the cluster that it belongs to versus the nearest cluster. It captures the intuition of how well-assigned the sample points in the clusters are. The range of the score varies between -1 to 1 with -1 indicating that a sample has been assigned to a wrong cluster and

Table 4.3: Internal evaluation of different clustering algorithms and distance metrics.

Algorithm	**Distance Measure**	**No. Clusters**	**CH**	**DB**	**SS**
k-means	ED	5	*10546.14*	*2.13*	0.35
k-means	DTW	5	9781.86	2.20	0.31
k-medoids	ED	5	8283.13	2.23	0.15
k-medoids	DTW	5	578.42	9.49	-0.12
k-shape	SBD	5	1568.62	5.57	-0.07
SOM	ED	5	9426.28	*2.13*	*0.40*

1 indicating that it has been assigned to the correct cluster. The silhouette score of an individual sample is

$$SS = \frac{\Delta' - \Delta}{max(\Delta, \Delta')}$$

where Δ' is the average distance between this point and points in the nearest cluster and Δ is the average distance between this point and points in the same cluster. The overall silhouette score is obtained by taking the average of the scores for all samples.

It can be seen from Table 4.3 that k-means with ED outperforms other methods considering DB and CH indices. Interestingly, SOM is on par with k-means in terms of the DB index, and outperforms it and all other methods in terms of SS.

4.2.2 Cluster Stability

Cluster stability is often used for model selection with the intuition that a clustering method should find similar clusters even if it is performed on different samples drawn from the same population [16]. However, we use this intuition to check the stability of the clusters over time. The basic idea is to check whether customers will stay in the same cluster or move to a different cluster if we run clustering at different points in time. The personas may need to be updated dynamically if a significant fraction of customers move to a different cluster over time. From the application perspective, the clusters need to be

stable to some extent over a period of time. Otherwise, by the time a product or strategy is developed and deployed, the segments will become invalid.

We evaluate the stability of the clusters using Jaccard Index (JI). Jaccard index (aka intersection over union) can be used to find out the similarity between two sets. If C_A and C_B are two sets it can be calculated as:

$$\text{JI}(C_A, C_B) = \frac{C_A \cap C_B}{C_A \cup C_B}$$

We compare our reference clusters of January 2019 with clusters obtained in February 2019 and in January 2020. As some of the customers are not present in the dataset for the latter two months, we only consider the customers that are available in both during the calculation. As an example, we calculate the Jaccard index for the first cluster of January 2019 with each of the 5 clusters obtained for February 2019. We pair the cluster from February 2019 which has the highest Jaccard index with the first cluster of January 2019 as these clusters are most similar to each other. We repeat this process for other clusters of January 2019.

To obtain a single Jaccard index for each algorithm rather than one index for each cluster created by an algorithm, we take the approach outlined below. Suppose there are n customers in the dataset. After all the clusters are paired we define a vector A of size n where the ith element of this vector is 1 if the ith customer which belongs to a certain cluster in one time period belongs to the cluster that is paired with it in another time period. Otherwise, this element is set to 0. Let A_0 and A_1 denote the total number of elements of A that are 0 and 1, respectively, The overall Jaccard index is calculated as follows:

$$\text{JI} = \frac{A_1}{A_0 + A_1}$$

We observed k-means with ED and SOM are most stable compared to other clustering methods as can be seen in Table 4.4. However, comparing the last two columns of this table it cannot be concluded that clusters become more unstable as time progresses. Additionally, we checked the Jaccard index between k-means with ED and DTW for January, 2019 and found it to be 0.897 which shows very high overlap in the clusters that they obtain. Based

Table 4.4: Analysis of stability over time.

Algorithm	**Distance Measure**	**No. Clusters**	**JI - Jan '19 & Feb '19**	**JI - Jan '19 & Jan '20**
k-means	ED	5	0.710	*0.702*
k-means	DTW	5	0.705	0.699
k-medoids	ED	5	0.491	0.633
k-medoids	DTW	5	0.453	0.466
k-shape	SBD	5	0.498	0.483
SOM	ED	5	*0.737*	0.672

on the fact that DTW is computationally complex and gives quite similar clusters to k-means with ED, we exclude it for the next two datasets. We, therefore, consider clusters formed by k-means with ED and SOM in the next step to mine association rules.

4.2.3 Cluster Shapes

The average representative pattern of the best-performing algorithms, i.e., k-means with ED and SOM with ED are shown in Figure 4.2 and Figure 4.4 along with k-means with DTW in Figure 4.3. This average monthly transaction pattern is drawn with a 90% confidence interval using bootstrapping and as can be seen the mean is extremely compact. The vertical lines represent last day of the week (Sunday). Naturally, the transaction frequency drops at that time. In Figure 4.2, Cluster 2 and Cluster 4 seem to have identical shapes but with different scales. The dominant cluster has a lower frequency of transactions compared to others. In Figure 4.4, we see similar shapes to Figure 4.2 although the numbers of cluster members differ significantly. Using the Jaccard index we pair the clusters found by these two methods as follows: (1-4, 2-2, 3-5, 4-1, 5-3). The Jaccard index is 0.768. We saw that k-means with ED and DTW identify very similar clusters which was also evident during the rule mining phase later on. This falls in line with the previous study mentioned in the literature review that found that for large scale data ED and DTW performs similarly. This allows us to to exclude it from consideration during business

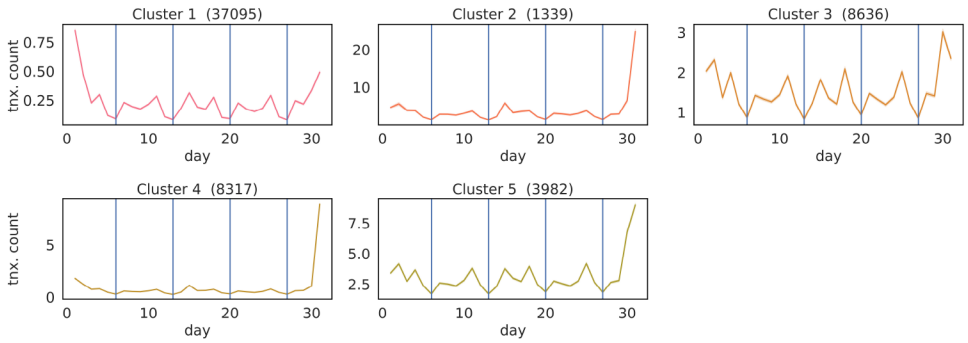

Figure 4.2: 5-means clustering with Euclidean distance (retail). The y-axis shows the number of transactions per day.

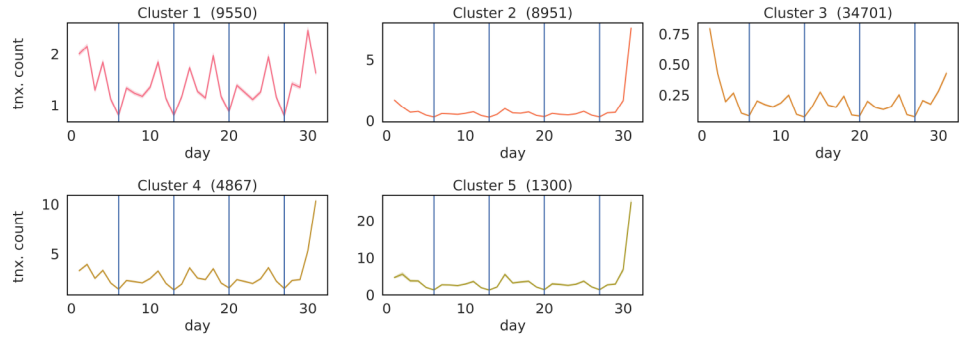

Figure 4.3: 5-means clustering with DTW distance (retail). The y-axis shows the number of transactions per day.

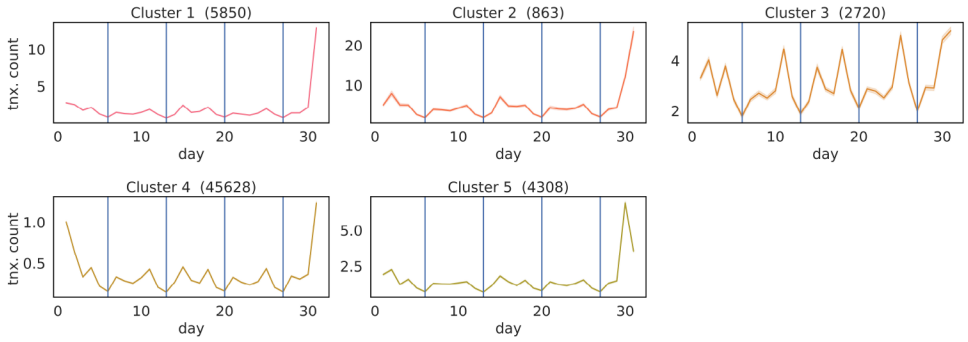

Figure 4.4: 5-cluster SOM clustering with Euclidean distance (retail). The y-axis shows the number of transactions per day.

customer and scalability study. It is can nonetheless be useful for cases where there is no way around clustering unequal length of transaction time series.

.

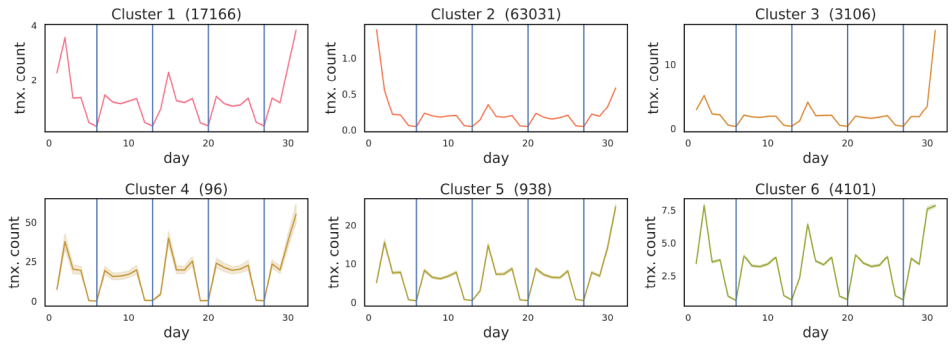

Figure 4.5: 6-means clustering with Euclidean distance (business). The y-axis shows the number of transactions per day.

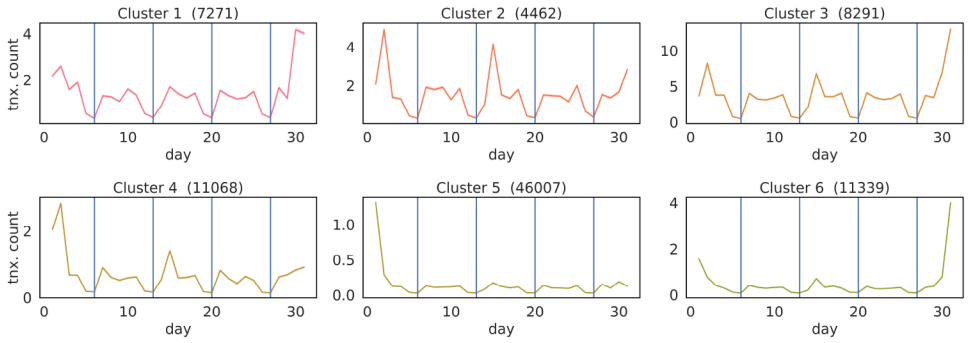

Figure 4.6: 6-cluster SOM clustering with Euclidean distance (business). The y-axis shows the number of transactions per day.

The transaction patterns for the business customers can be seen in Figure 4.5 and Figure 4.6 for k-means and SOM respectively. In this we also find a dominant cluster for both methods, although k-means one is more dominant is this case.

4.3 Analysis of Extracted Rules for Usefulness

Experts within the financial institution have seven personas that clients are categorized into based on demographic information such as age, net worth, and income. These personas were developed by analyzing the demographics of the client base and building narratives around the different types of clients that were observed. The examples of these personas can be seen in Figure 4.7 and Table 4.5.

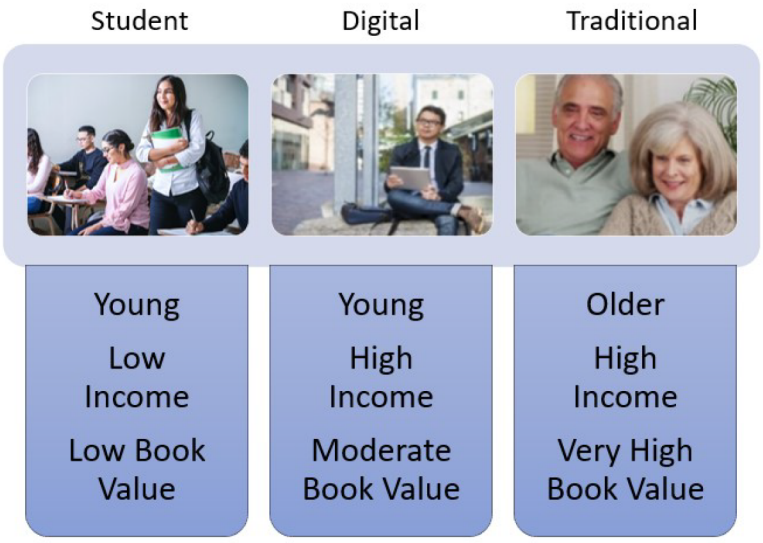

Figure 4.7: Examples of existing personas.

Using our systematic approach we extract a number of rules from the dataset. The number of rules obtained for retail and business customers can be seen in Table 4.6.

4.3.1 Retail Rules

The usefulness of the extracted rules can be qualitatively evaluated based on two aspects. One is if they are able to identify existing personas and the other is if they are able to discover new personas.

Table 4.5: Existing personas.

Persona Name	Age Bin	Income Bin	Book Value Bin
Student	1	1	1
Edge	1	3	1
Digital	1	3	2
Planner	2	3	3
Adventure	3	4	4
Traditional	4	3	4
Corporate	N/A	outlier	outlier

Table 4.6: Numbers of rules extracted.

Algorithm	Cluster	Size	Percentage	no. Rule Type 1	no. Rule Type 2
Retail k-means	1	37,095	62.80	0	390
	2	1,339	2.30	5	2
	3	8,636	14.60	1	539
	4	8,317	14.10	1	86
	5	3,982	6.70	2	1
Retail SOM	1	5,850	0.10	3	23
	2	863	0.01	3	0
	3	2,720	0.05	2	0
	4	45,628	0.77	0	847
	5	4,308	0.07	1	0
Business k-means	1	17,166	19.50	1	4
	2	63,031	71.80	0	126
	3	3,106	3.50	5	0
	4	96	0.10	1	0
	5	938	1.10	13	0
	6	4,101	4.70	3	3
Business SOM	1	7,271	8.30	1	0
	2	4,462	5.10	1	0
	3	8,291	9.40	5	132
	4	11,068	12.60	0	0
	5	46,007	52.40	0	32
	6	11,339	12.90	0	16

Alignment with Existing Personas

Cluster 3 of SOM corresponds to more transactions than a typical customer, but still relatively few. The transactions are highly cyclical week to week. In this cluster, the association rules found a grouping of young customers with low book value and moderate income. This aligns particularly well with the "Edge" persona.

SOM cluster 2, however, is defined by very large far more frequent transactions, and association rules mining finds that individuals in this cluster have very large income and book value. These are high value "corporate" clients that transact similarly to business customers. They are one of the key client

groups generating revenue for the financial institution. We find that cluster 2 of k-means also shows the same result.

Cluster 4 of k-means show a persona not found in the SOM clusters. These are people aged 48 to 62 with a large book value, a group that closely aligns with the "adventure" persona. These clusters are transactionally similar as well, with generally low transaction counts on most days except for at the end of the month.

Discovery of New Personas

SOM cluster 5 is one of the most interesting clusters because we find a rule that is not reflected in the predefined internal personas. The rule states that individuals with moderate income and low book value between 48 and 62 align with this cluster. There is an implicit assumption that as clients get older, they will have larger and larger savings (book value), but this is not observed.

Because every 5-cluster method grouped greater than 50% of all customers into a single cluster, which is consistently defined by a lower expected transaction count with only one major peak per month, this is a particularly critical cluster to look at. The SOM association rules mining results for this cluster show that 80% of all individuals over 62 years are in this cluster. 80% of people with low income and 80% of people with few products are here as well. There is a group of elderly, low income people with few accounts within this cluster that hardly transacts. This is a group of people that are not addressed by the existing personas. Additionally, there is a bias in the book value and income labels. A client that does not use the financial institution as their main one will appear to have low book value, low income, and few products. This is an association rule found with all 5-cluster methods for the large cluster. This is bolstered by the transactions in the cluster, which are very rare. The largest group of individuals at the institution are defined by their lack of transaction with the institution.

We found that we could use association rules mining and search for individuals with no gender and very high book value or income, a rule that showed up frequently in all clustering methods and was consistently aligned with clusters

that had high transaction frequency. We believe these to be business accounts that are mislabeled. That being said, there are individuals where gender is assigned that also meet the other criteria and are clustered with the high transacting business accounts. For these individuals, it is likely that some of these are also personal accounts that should be switched to business accounts.

4.3.2 Business Rules

The obtained business clusters have a similar distribution of clusters, with single dominant clusters of ultra-low transaction individuals. All other clusters have moderate to high transaction frequencies. This finding is validated by the rules mining, which shows that businesses with a lower income and book value are highly likely to be in the large single cluster.

Because the business rules can be associated with the NAICS code that aligns businesses to industry segments, they provide the financial institution with an industry-based split of high-value and low-value clients. Industries found in rules for high-value clients are: Retail Trade, Real Estate, and Arts and Entertainment. SOM rules mining does a better job of finding industry-based associations than k-means rules mining.

Industries found in rules for low-value clients are: Accommodation and Food Services, Management of Companies, Educational Services, and Public Administration. There are two explanations for why these customers are low-value. The first is that they are very small companies that are potentially struggling to do business. The second is that they do not use this financial institution as their primary one. In either case, the financial institution has been successful in attracting businesses from these industries, but not successful in converting their initial engagement into high value.

Based on the results of the association rules mining, we slightly prefer SOM to k-means. SOM, both in retail and business contexts, more consistently clusters customers in a way that leads to action that the financial institution can take.

Chapter 5

Performance Evaluation

5.1 Performance Analysis

The proposed framework must be scalable given the amount of data modern financial institutions need to handle. Our implementation in a real financial institution's cloud system gives us the unique capability to experiment within a practical scope. The choice of vertical and horizontal scalability depends on an organization's specific need in terms of dataset volume, ease of implementation, cost, and resource availability. We analyze vertical scalability within GCP with real data. On the other hand, we analyze horizontal scalability in external cloud platform using synthetic data. As the platform and underlying data are different, we do not compare them with each other. We rather show that the framework can be generalized in terms of scalability.

5.2 Distributed Implementation of the Framework for Scalability

We use Spark for distributed implementation. In contrast to Hadoop, Spark provides in-memory computation that makes it run faster. As many ML algorithms are iterative, Spark's in-memory computation is more suitable for such an application. We use distributed implementations of hierarchical, k-means, SOM, and association rule mining from MLib and other open-source libraries.

We use a variant of hierarchical clustering, bisecting k-means. It takes a divisive hierarchical clustering approach. Although in traditional hierarchical

clustering the number of clusters is not pre-specified, in bisecting k-means it is pre-specified. In the beginning, all the points of the dataset are contained in a single cluster. Then the algorithm finds divisible clusters on the bottom level and bisects each of them using k-means until there are k leaf clusters in total or no leaf clusters are divisible. The bisecting steps of clusters on the same level are grouped together to increase parallelism. If bisecting all the clusters at the bottom level generates more clusters than intended, then the larger clusters are bisected. For the other algorithms, the Spark implementation is the parallelized version of the main algorithm.

5.3 Vertical Scalability

Vertical scalability refers to adding more resources to a single node, usually through the addition of more Central Processing Unit (CPU), Random-Access Memory (RAM), or storage. We use the previous non-parallelized implementation within the GCP Platform for vertical scalability. We run all three modules of the pipeline, namely anomaly detection, clustering (both SOM and k-means for comparison), and rule mining for different customer data sizes up to 300,000 customers. It can be readily seen from Figure 5.1 that the runtime shows logarithmic growth on this plot which has a log scale Y-axis; this indicates that it scales linearly with the number of customers. In this experiment, we vertically scale our base Virtual Machine (VM) as needed. The workloads are RAM intensive. More specifically, in the anomaly detection and rule mining stage, we add more RAM to our base VM. Our base virtual machine has 4 vCPU and 26 GB RAM. Using extrapolation, a back-of-the-envelope calculation shows that running the whole pipeline will take approximately 137.62 minutes for 1 million customers.

5.4 Horizontal Scalability

Horizontal scalability refers to increasing the capacity of the system by adding more nodes to it. For the institution in question, vertical scalability sufficed. For horizontal scalability, we run the experiment in a different cloud envi-

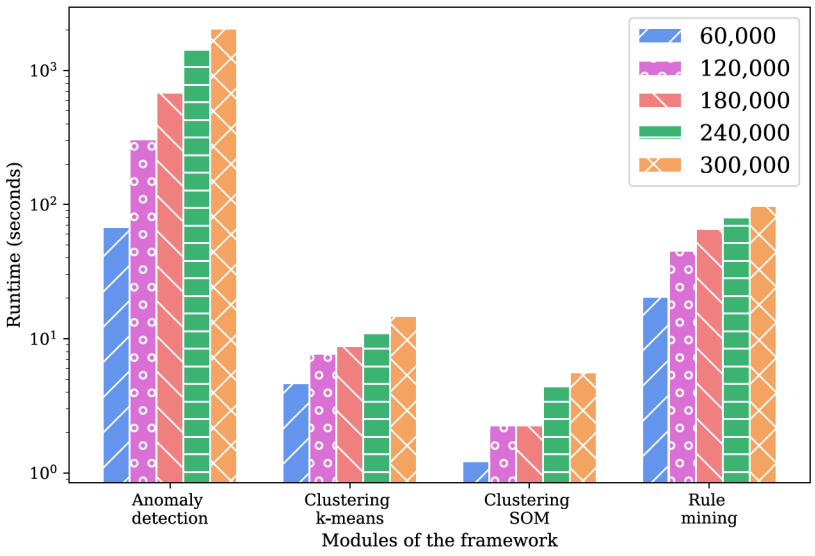

Figure 5.1: Runtime of each module of the pipeline for different number of customers. Note that the y-axis is in log scale.

ronment instead of GCP. As the real data is confidential, we use synthetic data instead. Generative models like Autoencoder, Generative Adversarial Network (GAN) and their variants are becoming popular for synthetic data generation [7], [103]. We use 500 transaction frequency time series for one month without any customer identification to generate 1 million synthetic data points using a Variational Autoencoder (VAE).

VAE [61] is a generative model that marries probabilistic graphical models (Bayesian networks) and deep learning. It generates latent variables from the input data and then generates synthetic data from the conditional distribution of the data given the latent variable. It uses gradient descent to minimize a loss function comprised of reconstruction loss and Kullback–Leibler (KL) divergence loss. The reconstruction loss is calculated in terms of error between the input and the generated sample. The KL divergence term is between the approximate posterior and true prior to the latent variable. Suppose X is the input data, $P(X)$ is the probability distribution of the data, z is the latent variable, $P(z)$ is the probability distribution of latent variable, and $P(X|z)$ is the distribution of generating data given latent variable. The final form of the loss function to minimize (provided that the prior and approximate posterior

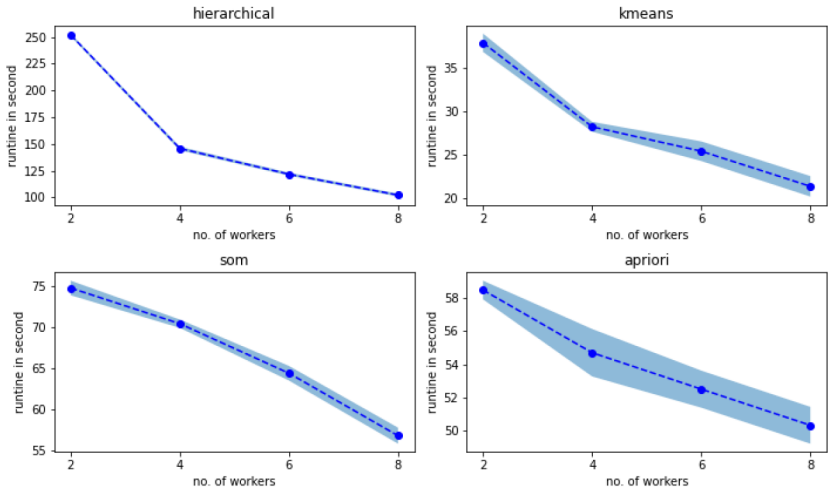

Figure 5.2: Runtime of each module of the pipeline for 1 million customers in different node configuration.

are multivariate Gaussian) is then given by [103]:

$$D_{KL}[N(\mu(X), \Sigma(X))\|N(0,1)] - \log P(X|z)$$

In our implementation, we pass the number of transactions made by a customer in a period of 31 days as a vector to the encoder, which maps it to a 4-dimensional latent space. The latent representation is the input to the decoder which aims to reconstruct the original data.

We run hierarchical, k-means, and SOM using this synthetic dataset. For association rule mining we use random data for 1 million customers. We use a virtual machine with 16 vCPU and 60 GB RAM. For the experiment, the driver has 4 vCPU and 8 GB RAM and each of the workers has 1 vCPU and 6 GB RAM. We experimented with 2, 4, 6, and 8 workers. We deploy different configurations of master and workers using docker. The master node processes the input and distributes the workload to the workers nodes. The runtime for the experiments can be seen in Figure 5.2. We exclude the first run for each module and take the average of the next 10 runs. The shaded area represents the standard deviation. As we see from the plot, the framework is linearly scalable.

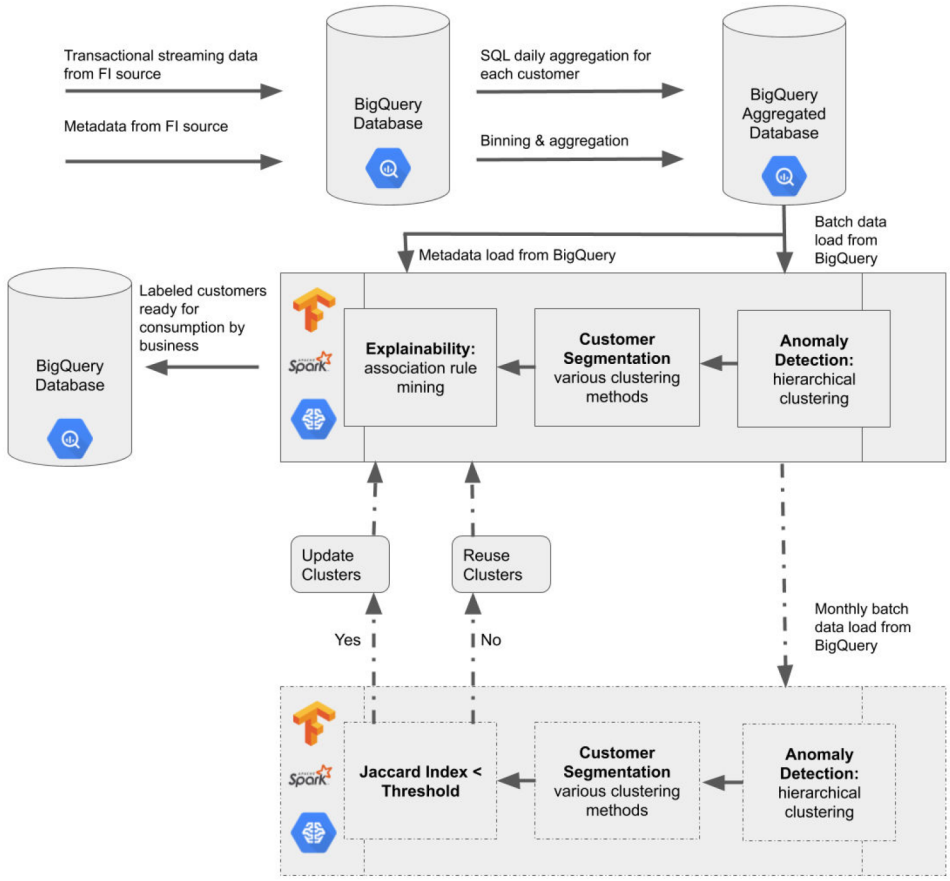

Figure 5.3: Architecture of the big data analytics pipeline in production.

5.5 Dynamic Nature

The dynamic nature of the framework comes from the fact that the clusters are automatically updated. The input data is streaming which is stored in BigQuery for a month. Then the framework starts to produce the clusters and provide them for business consumption. For the subsequent months, at the end of each month, the clusters are recomputed and Jaccard indexes between the clusters of two months are checked. If it falls below a user-defined threshold then the new clusters are set as default. This value can be tuned based on the requirement of stability. Figure 5.3 shows the dynamic distributed system in production.

Chapter 6

Conclusion

This thesis proposed an interpretable dynamic customer segmentation framework for financial institutions using anomaly detection, clustering, and association rule mining techniques in a practical manner. We have found compelling and actionable results in this paper by analyzing the rich dataset obtained from a financial institution.

6.1 Summary of Contributions

On the retail side, we have found several rules that are in perfect alignment with traditional personas of the financial institution, which is a good sign of internal validation of this work. More importantly, we have found rules that are not in alignment with any of the predefined personas. In particular, there is a group of elderly, low-income people with few accounts that hardly transact yet there is no persona that aligns with these individuals. Since these rules go against conventional banking wisdom that as clients get older, they will tend to have larger savings portfolios, this represents new opportunities for the bank to re-target its customers closer in actual alignment to what the data dictates. Additionally, we found that there is a large portion of customers who likely have a different financial institution as their primary one. We have also found a large portion of personal accounts that transact similarly to business accounts. Both of these findings represent unique opportunities for the bank to target existing customers and market products that better align to their transaction patterns.

On the business side, utilizing the NAICS codes and our rules mining approach has helped us find significant insights for tiering customers into high-value and low-value clients respectively, divided into useful industries to target for the bank. The single dominant cluster of ultra-low transaction individuals also appeared within the business clientele. As we elaborated above, this finding is particularly significant because the bank has been successful in attracting these businesses at large volumes, yet fails to capitalize on them to a profitable degree.

In summary, with our systematic association rules mining approach, we have also validated the existence of some of the personas developed by the financial institution while also finding new classes that are not acknowledged using traditional means. Our approach provides a method to keep personas aligned to the real distribution of demographics. By recalculating the clusters and association rules, we can validate the existence of customer groups not just through their demographics but also through their transaction patterns.

From an implementation standpoint, we have found that the algorithms we prefer to use for segmentation scale linearly with data. Additionally, we have found that the framework is vertically scalable if needed. Given that the financial transaction time series are rather stable, they can be reevaluated less frequently.

We believe it is important here to address the ethics of our customer segmentation approach. The primary purpose of this project is to understand the customer base as it is. If biases exist, it is essential for us to be able to capture them. Rather than feeding the results of the segmentation directly into another model without human intervention, this work can used by the bank to identify certain biases. The financial institution which provided us with data believes strongly that the best approach to eliminating bias is to build models with all available information and then compare them to demographic features such as age and gender to capture the biases that the model (and therefore the data) have. This is why the association rules mining step is critical to the ethical validity of this project.

6.2 Limitations

The study presented in this thesis has the following limitations:

1. We do not consider new customers up until we have at least 1 month worth of data. If there is a need for immediate inclusion of these new customers then the clustering problem can be turned into a classification problem. The new customers can be assigned to closest cluster center using DTW distance measure until 1 month worth of data is available.

2. It was hard to have direct external validation of the obtained clusters through domain experts.

3. We could not evaluate whether the obtained new segments were able to generate revenue based on appropriate application as this is a long-term process.

6.3 Future Directions

Several avenues can be pursued for future work following the study in this thesis.

1. We consider whole time series anomaly in our anomaly detection stage. There is an opportunity to explore windowed anomaly detection.

2. Given the application scenarios we converted the streaming data into batch data of 1 month. There may be specific applications that may require segmentation with streaming data. In such a case the work can be expanded with spark streaming.

3. Studying the effect of stability on a granular level to add meaning to the movement of individuals across clusters has significant value. It may also be possible to predict this movement from one cluster to another cluster and possibly one persona to another persona during a customer's lifetime.

4. We did not consider the cost aspect of horizontal vs vertical scalability in terms of cost, considering both cloud resources and personnel resources. This may be explored in a future study.

References

[1] M. R. Aburrous, A. Hossain, K. Dahal, and F. Thabatah, "Modelling intelligent phishing detection system for e-banking using fuzzy data mining," in *2009 International Conference on CyberWorlds*, IEEE, 2009, pp. 265–272.

[2] V. Aggelis and D. Christodoulakis, "Customer clustering using rfm analysis," in *Proceedings of the 9th WSEAS International Conference on Computers*, Citeseer, 2005, p. 2.

[3] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 1993, pp. 207–216.

[4] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.

[5] N. Albayrak, A. Özdemir, and E. Zeydan, "An overview of artificial intelligence based chatbots and an example chatbot application," in *2018 26th Signal processing and communications applications conference (SIU)*, IEEE, 2018, pp. 1–4.

[6] M. Alborzi and M. Khanbabaei, "Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed rfm analysis method," *International Journal of Business Information Systems*, vol. 23, no. 1, pp. 1–22, 2016.

[7] F. Alharbi, L. Ouarbya, and J. A. Ward, "Synthetic sensor data for human activity recognition," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–9.

[8] Ö. G. Ali and U. Arıtürk, "Dynamic churn prediction framework with more effective use of rare event data: The case of private banking," *Expert Systems with Applications*, vol. 41, no. 17, pp. 7889–7903, 2014.

[9] V. Andreo, X. Porcasi, C. Rodriguez, L. Lopez, C. Guzman, and C. M. Scavuzzo, "Time series clustering applied to eco-epidemiology: The case of aedes aegypti in córdoba, argentina," in *2019 XVIII Workshop on Information Processing and Control (RPIC)*, IEEE, 2019, pp. 93–98.

[10] A. S. Aribowo and N. H. Cahyana, "Feasibility study for banking loan using association rule mining classifier," *International Journal of Advances in Intelligent Informatics*, vol. 1, no. 1, pp. 41–47, 2015.

[11] W.-H. Au and K. C. Chan, "Mining fuzzy association rules in a bank-account database," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 2, pp. 238–248, 2003.

[12] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *2017 International Conference on Computing Networking and Informatics (ICCNI)*, IEEE, 2017, pp. 1–9.

[13] A. J. Bagnall and G. J. Janacek, "Clustering time series from arma models with clipped data," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 49–58.

[14] F. Batrinu, G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Efficient iterative refinement clustering for electricity customer classification," in *2005 IEEE Russia Power Tech*, IEEE, 2005, pp. 1–7.

[15] BCG, *What Does Personalization in Banking Really Mean?* `https://www.bcg.com/publications/2019/what-does-personalization-banking-really-mean`, Online; accessed on 21 September 2020, 2020.

[16] A. Ben-Hur *et al.*, "A stability based method for discovering structure in clustered data," in *Pacific Symposium on Biocomputing*, 2002, pp. 6–17.

[17] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series.," in *KDD workshop*, Seattle, WA, vol. 10, 1994, pp. 359–370.

[18] Business Insider, *AI in Banking*, `https://www.opentext.com/info/ai-financial-services`, Online; accessed on 21 September 2020, 2020.

[19] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[20] C. C. H. Chan, "Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer," *Expert systems with applications*, vol. 34, no. 4, pp. 2754–2762, 2008.

[21] Y. Chen, X. Liu, X. Li, X. Liu, Y. Yao, G. Hu, X. Xu, and F. Pei, "Delineating urban functional areas with building-level social media data: A dynamic time warping (dtw) distance based k-medoids method," *Landscape and Urban Planning*, vol. 160, pp. 48–60, 2017.

[22] G. Chicco and I.-S. Ilie, "Support vector clustering of electrical load pattern data," *IEEE Transactions on Power Systems*, vol. 24, no. 3, pp. 1619–1628, 2009.

[23] G. Chicco, O.-M. Ionel, and R. Porumb, "Electrical load pattern grouping based on centroid model with ant colony clustering," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1706–1715, 2012.

[24] G. Chicco, R. Napoli, and F. Piglione, "Load pattern clustering for short-term load forecasting of anomalous days," in *2001 IEEE Porto Power Tech Proceedings (Cat. No. 01EX502)*, IEEE, vol. 2, 2001, 6–pp.

[25] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Emergent electricity customer classification," *IEE Proceedings-Generation, Transmission and Distribution*, vol. 152, no. 2, pp. 164–172, 2005.

[26] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," *IEEE Transactions on Power Systems*, vol. 18, no. 1, pp. 381–387, 2003.

[27] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. Kurgan, *Data mining: a knowledge discovery approach*. Springer Science & Business Media, 2007.

[28] L. Columbus, "Why ai is the future of financial services," 2019.

[29] Data Axle, *The finance industry's guide to marketing data*, `https://www.data-axle.com/resource/the-finance-industrys-guide-to-marketing-data/?sfcid=7010d000001L1JG`, Online; accessed on 21 September 2020, 2020.

[30] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[31] I. Dent, T. Craig, U. Aickelin, and T. Rodden, "An approach for assessing clustering of households by electricity usage," *Available at SSRN 2828465*, 2012.

[32] S. Dhankhad, E. Mohammed, and B. Far, "Supervised machine learning algorithms for credit card fraudulent transaction detection: A comparative study," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, IEEE, 2018, pp. 122–125.

[33] N. G. Dincer and Ö. Akkuş, "A new fuzzy time series model based on robust clustering for forecasting of air pollution," *Ecological Informatics*, vol. 43, pp. 157–164, 2018.

[34] R. Ding, Q. Wang, Y. Dang, Q. Fu, H. Zhang, and D. Zhang, "Yading: Fast clustering of large-scale time series data," *Proceedings of the VLDB Endowment*, vol. 8, no. 5, pp. 473–484, 2015.

[35] M. Disegna, P. D'Urso, and F. Durante, "Copula-based fuzzy clustering of spatial time series," *Spatial Statistics*, vol. 21, pp. 209–225, 2017.

[36] V. Estivill-Castro and J. Yang, "Fast and robust general purpose clustering algorithms," in *Pacific Rim International Conference on Artificial Intelligence*, Springer, 2000, pp. 208–218.

[37] EY, *Future Consumer Index: How COVID-19 is changing consumer behaviors*, `https://www.ey.com/en_ca/consumer-products-retail/how-covid-19-could-change-consumer-behavior`, Online; accessed on 21 September 2020, 2020.

[38] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Transactions on power systems*, vol. 20, no. 2, pp. 596–602, 2005.

[39] Forbes, *How AI Can Improve Financial Analytics*, `https://www.forbes.com/sites/louiscolumbus/2020/07/23/how-ai-can-improve-financial-analytics`, Online; accessed on 21 September 2020, 2020.

[40] D. Gallego and G. Huecas, "An empirical case of a context-aware mobile recommender system in a banking environment," in *2012 third FTRA international conference on mobile, ubiquitous, and intelligent computing*, IEEE, 2012, pp. 13–20.

[41] M. Gavrilas, G. Gavrilas, and C. V. Sfintes, "Application of honey bee mating optimization algorithm to load profile clustering," in *2010 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*, IEEE, 2010, pp. 113–118.

[42] D. Gerbec, S. Gašperič, I. Šmon, and F. Gubina, "Determining the load profiles of consumers based on fuzzy logic and probability neural networks," *IEE Proceedings-Generation, Transmission and Distribution*, vol. 151, no. 3, pp. 395–400, 2004.

[43] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Allocation of the load profiles to consumers using probabilistic neural networks," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 548–555, 2005.

[44] D. Gerbec, S. Gasperic, and F. Gubina, "Determination and allocation of typical load profiles to the eligible consumers," in *2003 IEEE Bologna Power Tech Conference Proceedings,*, IEEE, vol. 1, 2003, 5–pp.

[45] F. Gullo, G. Ponti, A. Tagarelli, S. liritano, M. Ruffolo, and D. Labate, "Low-voltage electricity customer profiling based on load data clustering," in *Proceedings of the 2009 International Database Engineering & Applications Symposium*, 2009, pp. 330–333.

[46] C. Guo, H. Jia, and N. Zhang, "Time series clustering based on ica for stock data analysis," in *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*, IEEE, 2008, pp. 1–4.

[47] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2013.

[48] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM sigmod record*, vol. 29, no. 2, pp. 1–12, 2000.

[49] T. S. Harrison, "Mapping customer segments for personal financial services," *International Journal of Bank Marketing*, vol. 12, no. 8, pp. 17–25, 1994.

[50] H. Hassani, X. Huang, and E. Silva, "Digitalisation and big data mining in banking," *Big Data and Cognitive Computing*, vol. 2, no. 3, p. 18, 2018.

[51] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining—a general survey and comparison," *ACM sigkdd explorations newsletter*, vol. 2, no. 1, pp. 58–64, 2000.

[52] G. T. Ho, W. Ip, C. Lee, and W. Mou, "Customer grouping for better resources allocation using ga based clustering technique," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1979–1987, 2012.

[53] N.-C. Hsieh, "An integrated data mining and behavioral scoring model for analyzing bank customers," *Expert systems with applications*, vol. 27, no. 4, pp. 623–633, 2004.

[54] X. Hu, H. L. Zhang, X. Wu, J. Chen, Y. Xiao, Y. Xue, T. Li, and H. Zhao, "A new customer segmentation framework based on biclustering analysis.," *JSW*, vol. 9, no. 6, pp. 1359–1366, 2014.

[55] A. M. Hughes, *Strategic database marketing: the masterplan for starting and managing a profitable, customer-based marketing program.* McGraw-Hill New York, NY, 2000, vol. 12.

[56] D. Jang, J. Eom, M. J. Park, and J. J. Rho, "Variability of electricity load patterns and its effect on demand response: A critical peak pricing experiment on korean commercial and industrial customers," *Energy Policy*, vol. 88, pp. 11–26, 2016.

[57] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[58] S. S. Jung and W. Chang, "Clustering stocks using partial correlation coefficients," *Physica A: Statistical Mechanics and its Applications*, vol. 462, pp. 410–420, 2016.

[59] J. Kang and J.-H. Lee, "Electricity customer clustering following experts' principle for demand response applications," *Energies*, vol. 8, no. 10, pp. 12 242–12 265, 2015.

[60] S.-Y. Kim, T.-S. Jung, E.-H. Suh, and H.-S. Hwang, "Customer segmentation and strategy development based on customer lifetime value: A case study," *Expert systems with applications*, vol. 31, no. 1, pp. 101–107, 2006.

[61] D. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[62] B. Knab, A. Schliep, B. Steckemetz, and B. Wichern, "Model-based clustering with hidden markov models and its application to financial time-series data," in *Between Data Science and Applied Data Analysis*, Springer, 2003, pp. 561–569.

[63] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[64] P. Kotler, *Marketing Management*. Prentice Hall International, 1994, vol. 8.

[65] R. Kuo, Y. An, H. Wang, and W. Chung, "Integration of self-organizing feature maps neural network and genetic k-means algorithm for market segmentation," *Expert systems with applications*, vol. 30, no. 2, pp. 313–324, 2006.

[66] S. Lahmiri, "Clustering of casablanca stock market based on hurst exponent estimates," *Physica A: Statistical Mechanics and its Applications*, vol. 456, pp. 310–318, 2016.

[67] R. Lamedica, L. Santolamazza, G. Fracassi, G. Martinelli, and A. Prudenzi, "A novel methodology based on clustering techniques for automatic processing of mv feeder daily load patterns," in *2000 Power Engineering Society Summer Meeting (Cat. No. 00CH37134)*, IEEE, vol. 1, 2000, pp. 96–101.

[68] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the hadoop ecosystem," *Journal of Big Data*, vol. 2, no. 1, p. 24, 2015.

[69] J. Lee, S. Yoo, H. Kim, and Y. Chung, "The spatial and temporal variation in passenger service rate and its impact on train dwell time: A time-series clustering approach using dynamic time warping," *International Journal of Sustainable Transportation*, vol. 12, no. 10, pp. 725–736, 2018.

[70] G. Lefait and T. Kechadi, "Customer segmentation architecture based on clustering techniques," in *2010 Fourth International Conference on Digital Society*, IEEE, 2010, pp. 243–248.

[71] T. Li, X. Wu, and J. Zhang, "Time series clustering model based on dtw for classifying car parks," *Algorithms*, vol. 13, no. 3, p. 57, 2020.

[72] T. W. Liao, "Clustering of time series data—a survey," *Pattern recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[73] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *Journal of the American statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[74] J. Lin, Z. Zhang, J. Zhou, X. Li, J. Fang, Y. Fang, Q. Yu, and Y. Qi, "NetDP: An industrial-scale distributed network representation framework for default prediction in ant credit pay," in *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, pp. 1960–1965.

[75] B. Liu, W. Hsu, Y. Ma, *et al.*, "Integrating classification and association rule mining.," in *KDD*, vol. 98, 1998, pp. 80–86.

[76] Q. Ma, J. Zheng, S. Li, and G. W. Cottrell, "Learning representations for time series clustering," in *Advances in neural information processing systems*, 2019, pp. 3781–3791.

[77] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, vol. 1, 1967, pp. 281–297.

[78] N. Mahmoudi-Kohan, M. P. Moghaddam, M. Sheikh-El-Eslami, and S. Bidaki, "Improving wfa k-means technique for demand response programs applications," in *2009 IEEE Power & Energy Society General Meeting*, IEEE, 2009, pp. 1–5.

[79] I. Maryani and D. Riana, "Clustering and profiling of customers using rfm for customer relationship management recommendations," in *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, IEEE, 2017, pp. 1–6.

[80] J. McWaters, "The new physics of financial services: Understanding how artificial intelligence is transforming the financial ecosystem," in *World Economic Forum*, 2018.

[81] J. H. Min and Y.-C. Lee, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters," *Expert systems with applications*, vol. 28, no. 4, pp. 603–614, 2005.

[82] A. Mutanen, M. Ruska, S. Repo, and P. Jarventausta, "Customer classification and load profiling method for distribution systems," *IEEE Transactions on Power Delivery*, vol. 26, no. 3, pp. 1755–1763, 2011.

[83] M. Namvar, M. R. Gholamian, and S. KhakAbi, "A two phase clustering method for intelligent customer segmentation," in *2010 International Conference on Intelligent Systems, Modelling and Simulation*, IEEE, 2010, pp. 215–219.

[84] J. Nazarko and Z. A. Styczynski, "Application of statistical and neural approaches to the daily load profiles modelling in power distribution systems," in *1999 IEEE transmission and distribution conference (Cat. No. 99CH36333)*, IEEE, vol. 1, 1999, pp. 320–325.

[85] A. Notaristefano, G. Chicco, and F. Piglione, "Data size reduction with symbolic aggregate approximation for electrical load pattern grouping," *IET Generation, Transmission & Distribution*, vol. 7, no. 2, pp. 108–117, 2013.

[86] C. Ozveren, C. Vechakanjana, and A. Birch, "Fuzzy classification of electrical load demand profiles-a case study," 2002.

[87] I. P. Panapakidis, T. A. Papadopoulos, G. C. Christoforidis, and G. K. Papagiannis, "Pattern recognition algorithms for electricity load curve analysis of buildings," *Energy and Buildings*, vol. 73, pp. 137–145, 2014.

[88] J. Paparrizos and L. Gravano, "K-shape: Efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015, pp. 1855–1870.

[89] Y. S. Patel, D. Agrawal, and L. S. Josyula, "The rfm-based ubiquitous framework for secure and efficient banking," in *2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH)*, IEEE, 2016, pp. 283–288.

[90] S. Patil, V. Nemade, and P. K. Soni, "Predictive modelling for credit card fraud detection using data analytics," *Procedia computer science*, vol. 132, pp. 385–395, 2018.

[91] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition*, vol. 44, no. 3, pp. 678–693, 2011.

[92] T. C. Phan, M. O. Rieger, and M. Wang, "Segmentation of financial clients by attitudes and behavior," *International Journal of Bank Marketing*, 2019.

[93] B. Pitt and D. Kitschen, "Application of data mining techniques to load profiling," in *Proceedings of the 21st International Conference on Power Industry Computer Applications. Connecting Utilities. PICA 99. To the Millennium and Beyond (Cat. No. 99CH36351)*, IEEE, 1999, pp. 131–136.

[94] T. Räsänen and M. Kolehmainen, "Feature-based clustering for electricity use time series data," in *International conference on adaptive and natural computing algorithms*, Springer, 2009, pp. 401–412.

[95] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[96] H. Sakoe, S. Chiba, A. Waibel, and K. Lee, "Dynamic programming algorithm optimization for spoken word recognition," *Readings in speech recognition*, vol. 159, p. 224, 1990.

[97] N. Sun, J. G. Morris, J. Xu, X. Zhu, and M. Xie, "iCARE: A framework for big data-based banking customer analytics," *IBM Journal of Research and Development*, vol. 58, no. 5/6, pp. 4–1, 2014.

[98] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1120–1128, 2007.

[99] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, 2015.

[100] A. Vellido, P. Lisboa, and K. Meehan, "Segmentation of the on-line shopping market using neural networks," *Expert systems with applications*, vol. 17, no. 4, pp. 303–314, 1999.

[101] A. Al-Wakeel, J. Wu, and N. Jenkins, "K-means based load estimation of domestic smart meter measurements," *Applied energy*, vol. 194, pp. 333–342, 2017.

[102] N. Waminee, S. Anongnart, and K. Sukumal, "Clustering ebanking customer using data mining and marketing segmentation," *ECTI Transactions on Computer and Information Technology*, vol. 2, no. 1, 2006.

[103] Z. Wan, Y. Zhang, and H. He, "Variational autoencoder based synthetic data generation for imbalanced learning," in *2017 IEEE symposium series on computational intelligence (SSCI)*, IEEE, 2017, pp. 1–7.

[104] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 275–309, 2013.

[105] W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," *World Wide Web*, vol. 16, no. 4, pp. 449–475, 2013.

[106] H. Weytjens, E. Lohmann, and M. Kleinsteuber, "Cash flow prediction: Mlp and lstm compared to arima and prophet," *Electronic Commerce Research*, pp. 1–21, 2019.

[107] T. Wittman, "Time-series clustering and association analysis of financial data," *University of Texas, Austin*, 2002.

[108] Y. Xiao, J. Yang, H. Que, M. J. Li, and Q. Gao, "Application of wavelet-based clustering approach to load profiling on ami measurements," in *2014 China International Conference on Electricity Distribution (CICED)*, IEEE, 2014, pp. 1537–1540.

[109] Y. Xie, X. Li, E. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5445–5449, 2009.

[110] N. Yamaguchi, J. Han, G. Ghatikar, S. Kiliccote, M. A. Piette, and H. Asano, "Regression models for demand reduction based on cluster analysis of load profiles," in *2009 IEEE PES/IAS Conference on Sustainable Alternative Energy (SAE)*, IEEE, 2009, pp. 1–7.

[111] J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S. E. Lee, C. Sekhar, and K. W. Tham, "K-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement," *Energy and Buildings*, vol. 146, pp. 27–37, 2017.

[112] X. Yang, J. Chen, P. Hao, and Y. J. Wang, "Application of clustering for customer segmentation in private banking," in *Seventh International Conference on Digital Image Processing (ICDIP 2015)*, International Society for Optics and Photonics, vol. 9631, 2015, 96311Z.

[113] W. Yotsawat and A. Srivihok, "Inbound tourists segmentation with combined algorithms using k-means and decision tree," in *The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, IEEE, 2013, pp. 189–194.

[114] D. Zakrzewska and J. Murlewski, "Clustering algorithms for bank customer segmentation," in *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, IEEE, 2005, pp. 197–202.

[115] H. Ziegler, M. Jenny, T. Gruse, and D. A. Keim, "Visual market sector analysis for financial time series data," in *2010 IEEE Symposium on Visual Analytics Science and Technology*, IEEE, 2010, pp. 83–90.