UNIVERSITY OF TWENTE.

Applications of Early Warning Systems for Customer Segmentation of Wholesale Banking Clients

by

Alessandra Amato

A thesis submitted to the
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
in partial fulfilment of the requirements for the degree of

MSc in Business Information Technology

Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)

University of Twente

Enschede, Overijssel, The Netherlands

October 2023

ABSTRACT

In the rapidly evolving landscape of Wholesale Banking (WB), Early Warning Systems (EWS) has increasingly become a vital resource for financial institutions aiming at monitoring their credit portfolio and preemptively detecting financial distress scenarios. For instance, ING Bank has tried to leverage the overwhelming wave of data imposed by the phenomenon of Big Data by implementing the Advanced Risk Integrated Application (ARIA), the company's EWS tool developed to surveil their commercial clients and to generate a number of warning in the presence of a potential risk incurring. However, since the current active triggers are only capable of detecting ongoing negative changes, ING has tried to explore innovative ways to expand the value delivered by the tool and introduce new solutions for the identification of potential up-selling opportunities. Among all the possible data-driven techniques that nowadays companies have started to rely on in order to maximise revenues and enhance their profitability, automated Customer Segmentation (CS) represents one the most successful and effective techniques developed. Therefore, the goal of this study focused on the investigation and implementation of a novel CS model, integrating early warning triggers, by answering the following main research question:

> *How to design and integrate early warning signals into a new CS model in order to identify potential business opportunities within banks' WB credit portfolio entities?*

In order to align the outcomes of the model developed to the initial business objectives, the research defined a number of requirements that the artifact should have presented related to its segments' orientation, identifiability and actionability. On the basis of the aforementioned characteristics, the research designed and introduced several different variables that aimed at providing a comprehensive and complete overview of the risk scenario associated with each client. The attributes in question, which can be obtained from the preprocessing and feature engineering of historical records of clients' internal data, internal triggers and external triggers, defined the client's risk profile from several perspectives: the progress and growth the entities have faced through the months in terms of EAD, RWA, allocated limit, outstanding amount and expected loss, the evolution of the client's credit quality rating, the average number of monthly early warning raised by each borrower and, finally, the client's current activity status, credit limit and outstanding balance recorded in the last month of the study.

On the basis of the insights derived from a systematic literature review on the application of EWS and CS in the field of finance, two popular clustering algorithms have been deployed, namely K-Means and DBSCAN, along with dimensionality reduction techniques such as correlation analysis and Principal Component Analysis (PCA). Moreover, the Elbow Method and

Silhouette Score were also used to validate the models deployed.

The assessment and interpretation of the clusters generated was performed through the implementation of a number of analyses that explored the different segments from multiple aspects, such as the tightness and separation of the subgroups formed or the density and descriptive statistics of the customers' distribution. From these studies it was discovered that the use of PCA slightly improved the compactness and distinction among the clusters, compared to the dataset derived from the correlation analysis. In addition, it was also observed that DBSCAN clustering algorithm proved to be unsuitable and inefficient for the type of data under examination, as no real conclusion and meaningful insight could be derived from the exploration of its clusters. Finally, a risk-reward analysis and risk exposure analysis related to, respectively, the comparison between the average number of negative and positive monthly triggers and the juxtaposition of the growths detected for the outstanding amount and the respective EAD value, were included as well.

In conclusion, the research contributed to obtaining a deeper understanding of the financial health of ING's WB clients, enabling decision-makers to re-adapt strategies and deliver more custom and targeted services based on each segment emerging needs. In addition, the study was able to bridge the gap between EWS and CS by introducing a novel perspective on strategic risk monitoring.

Keywords: Early Warning Systems; Customer Segmentation; Unsupervised Machine Learning; Wholesale Banking; Financial Industry; Lending

AUTHOR'S DECLARATION

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Twente to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Twente to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

Alessandra Amato

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the following individuals for their invaluable support throughout the course of this project:

Firstly, I want to express my most sincere appreciation to my thesis supervisors, Marcos Machado, Jörg Osterrieder and João Rebelo Moreira, for sharing their expertise and guiding me in the development of my thesis. In particular, I would like to thank Marcos for introducing me to the project, providing timely, comprehensive and consistent feedback, and for the moral encouragement given during the past year. In addition, I want to extend my gratitude to Anand Autar for allowing me to join ING and the ARIA team for my graduation internship experience.

To my colleagues and peers in ING. I am grateful to have had the chance to work with the ARIA squad, including Rui Santos, Mehmet Simsek, Peter Lichtenveldt, Christopher Pironti, Krzysztof Mirek, Wioleta Ranik, Michał Kajstura, Piotr Treska, and Daniel Chen. I would like to express my deepest recognition to them for warmly welcoming me into the team as a valuable component and generously sharing their time and knowledge with me during the course of my thesis work. Special thanks go to my fellow Italian teammate, Christopher, for his collaborative spirit and dedication in assisting me throughout the project's development with new concepts and ideas, and my nearly-Italian colleague, Robin Zijp, for connecting me with several interesting individuals within ING and for making sure that I had access to all the necessary resources. Moreover, I am immensely thankful to Rui, for giving me this opportunity and offering assistance whenever I found myself in difficulty.
I also would like to extend my appreciation to all the Data Science Chapter members. Their work and commitment represented a real source of inspiration that I believe will play a pivotal role in shaping my future career as well.

Lastly, I would like to express my deepest and most sincere gratitude to my entire family, without whom this thesis and experience would not have been possible. To my parents, Fede, Olivia and Benjamin, whose constant encouragement, understanding and love provided the necessary fuel for my determination and motivation, despite my occasional bad temper. Furthermore, my heartfelt appreciation goes to my friends from both my master's programme and my hometown in Trieste. In particular, I want to deliver special and important thanks to my dearest friend, Federica, for being my steadfast companion and helping me maintain my sanity over the last ten years, and to Suraj, for patiently supporting me during the challenging times.

Thank you all for your contribution to this thesis and for guiding me toward the accomplishment of this significant milestone in my life. I hope you enjoy the reading!

CONTENTS

List of Figures

LIST OF TABLES

ABBREVIATIONS

ARIA	Advanced Risk Integrated Application.
CDS	Credit Default Swap.
CRISP-DM	CRoss Industry Standard Process for Data Mining.
CS	Customer Segmentation.
DSR	Design Science Research.
EAD	Exposure At Default.
EWS	Early Warning Systems.
LGD	Loss Given Default.
ML	Machine Learning.
NLP	Natural Language Processing.
PCA	Principal Component Analysis.
PCs	Principal Components.
PD	Probability of Default.
RB	Retail Banking.
RWA	Risk-Weighted Assets.
WB	Wholesale Banking.
WCSS	Within Cluster Sum of Squares.

1

INTRODUCTION

In today's highly demanding lending market, financial institutions are constantly in search for innovative solutions that would allow them to stay ahead the competition and enhance the business profitability. With the emergence of the phenomenon of Big Data and the development of advanced analytic techniques, allowing experts to explore and manage data from large and diverse datasets, organisations are now capable of acquiring more insightful information related to their customers and developing more targeted strategies [3]. Among all the different data-driven implementations that companies have integrated in their workflow, one of the most popular and efficient techniques is Customer Segmentation (CS). The term refers to the process of dividing a large base of clients into smaller subgroups that share similar characteristics and behaviours relevant to the marketing and sales goals of the organisation [4].

With regards to the financial sector and the lending field, in particular, this technology helps banks and institutions to offer more tailored products and services based on the needs that each segment of customers manifests and improve the risk management procedures as well [5]. Moreover, by putting in place a CS, or customer clustering, application marketers can also identify cross and up-selling opportunities at a glance. However, the performance of the segmentation developed and its efficiency for the business highly depend on the type of features taken into consideration to build the respective clusters, reflecting specific behaviours and traits of the clients [6]. Therefore, the selection of the pertinent and significant features represents a critical step that not only requires an in-depth understanding of the business and its target audience but that also ought to be consistent with the initial business objective that led to the development of the application [7].

Another state-of-the-art implementation that has seen a significant growth during the last decade and has now become a valuable asset for credit risk monitoring are Early Warning Systems (EWS). EWS can be defined as qualitative and quantitative indicators capable of anticipating risk events [8]. The generation of one signal, also known as trigger, preemptively informs the

management board of a future distress situation and allows them to design the appropriate mitigation measures and follow-up actions.
Although EWS still represent an unexplored field for most enterprises, more and more organisations have started to acknowledge their power and introduce them in their current risk management practices, especially in the financial sector [9]. For this reason, the integration of the information obtained from the use of EWS within CS systems may represent a valuable element that could further improve the outcome of these processes. The inclusion of the risk factor involved when engaging with insolvent clients would, indeed, provide insights on the current and future credit health of the portfolio entities of each segment.

1.1. RESEARCH BACKGROUND

ING is a multinational financial institution established in the Netherlands with more than 60,000 employees, providing banking, insurance and assets management services to both Retail Banking (RB) and Wholesale Banking (WB) entities [10, 11]. Among all the different products that ING currently offers, lending represents one of the main services that the bank grants in order to support its WB customers, including large corporations, governments or other financial institutions as well. Compared to RB, WB deals with financial transactions of a larger scale and is highly vulnerable to all the macroeconomic trends that shape the market landscape and drive its transformation [12]. For example, due to the Covid-19 pandemic, in 2020 the world-wide WB sector faced a significant increase of uncertainty and risk which deeply affected banks' lending decision and led to the development of new strategies and programs to support businesses [13]. Therefore, WB tends to be exposed to a greater range of risks and threats, deriving from the type of operations that the involved institutions perform and the market changes. Given the considerable amount of capital that has to be administered, the effective monitoring of the loans and clients' repayment capabilities represents a paramount lifeguard for banks and institutions.
As one of the biggest lending institutions in the world, ING has tried to address this issue and to overcome the challenge imposed by the phenomenon of Big Data, generating a tsunami of information that can hinder the identification of relevant news, by developing an innovative digital warning system. The system, known as the Advanced Risk Integrated Application (ARIA), aims to provide timely and actionable early warning signals related to a number of risk-related indicators, such as entities' Probability of Default (PD) or Exposure At Default (EAD), which would allow risk managers and front officers to take preventative measures and mitigate the potential impacts that these events would have on the bank's business.
In order to implement the ARIA application, data is collected from both internal and external data sources. The data concerning specifically client's personal information is retrieved from existing internal systems and infrastructures of ING and stored into one of ARIA's central databases. External data sources, instead, are deployed to collect market indicators, for example stock prices, and online news articles involving the credit portfolio entities that must be monitored. Once all the data is fetched, the system is able to generate an early warning trigger if

one of the indicators in question presents a considerable change over a certain period time. For instance, if application detects a decrease of Equity of more than 10% on a daily basis for one particular entity, a trigger will be raised for the respective client, signaling a potential increase of risk. As previously mentioned, ARIA is also capable discovering online articles, available on external information sources such as Google News, Baisu News and the Financial Times, concerning a specific topic and involving ING's WB institutions. This functionality makes use of Natural Language Processing (NLP) techniques for entity recognition and content classification, and clustering algorithms to group news articles into events.
The solution offered by ING and the ARIA team represents a fundamental resource for the company, as not only it allows to obtain a more accurate and broader picture on the financial health of the client base, but also to ensure more stability and resilience within the credit portfolio.

1.2. RESEARCH MOTIVATIONS AND OBJECTIVES

Over the last few years, ARIA has significantly reduced the workload of risk managers and accelerated their decision-making processes by ensuring a continuous monitoring of markets, sectors and organisations in a proactive manner. However, regardless of these notable successes, it is important to underline that the early warning signals currently raised by the tool only aim at discovering unexpected negative changes and clients in distress. In fact, it can be affirmed that, as of today, the system is mainly risk-oriented and is not yet capable of generating warnings in the case of, for example, prospective business opportunities. Therefore, the current value delivered by the application could be further extended by introducing new analytical solutions which would advice front office managers of the existence of potential up-selling and business-making scenarios.
Nowadays, several data-driven techniques can be utilised by financial institutions to maximise revenues and deepen the relationships with customers. Automated CS represents one of the most efficient and successful solutions that most companies have started to adopt to develop more targeted business initiatives.
As ING's main strategy is founded on the principle of customer empowerment with the WB objective to become more client-centric to stay ahead of evolving trends, the bank has already established a global client segmentation model that allows them to determine the type of coverage and service levels each segment requires. The segments created, however, are solely based on financial and behavioural historical records of the entities. Given ARIA's power to provide detailed and timely insights on the current financial health conditions of WB clients, it can be argued that the integration of the information released by the application into a new CS model may lead to the generation of even more accurate and informative segments, which would enable the identification of low-risk prospects for further business-making opportunities.

The primary goal of this study, therefore, is to make use of the triggers raised by an EWS tool to design a new automated CS model using unsupervised Machine Learning (ML) algorithms. Based on the data available in ARIA's databases, the research investigates the appropriate fea-

tures that can be considered to build the segments, or clusters, of similar entities. In addition, the research focuses on the analysis of the results obtained from different clustering models to identify potential groups of clients which present ideal entity profiles for credit extensions and up-selling opportunities. The scope of this project's use case, however, is only limited to the WB credit portfolio entities that are currently being monitored by the application.
With the deployment of this model, several advantages could be achieved:

1. More targeted strategies: by dividing customers into distinct segments, sharing similar characteristics and behaviours, stakeholders would be able to define appropriate actions for each group and improve the resource allocation within the credit portfolio.
2. Swift identification of business opportunities: the insights derived from the segmentation would enhance customer understanding which, consequently, may also lead to a swifter and rapid identification of possible advantageous opportunities.
3. Increased customer retention: the understanding of customers segments would also allow the bank to address the needs manifested by each group and offer a more proactive customer service. This way the bank would be able to increase customers' satisfaction and loyalty.

Based on the above-mentioned research objectives, the study aims at providing answers to the following main research question:

> *How to design and integrate early warning signals into a new CS model in order to identify potential business opportunities within banks' WB credit portfolio entities?*

From the main research questions, several other sub-questions have been outlined:

1. How can the information obtained from the EWS be processed in order to be integrated in the CS model?
   (a) What requirements should such model present?
   (b) What type of variables can be deployed to generate informative customer clusters that are aligned to the initial business objectives?
2. How can customer segments be generated in an automated manner?
   (a) What ML techniques can be used?
   (b) How can the decision-making processes of the algorithms used be explained?
3. How can the quality of the clusters generated be assessed and how can the segments obtained be interpreted?
   (a) What type of clustering evaluation metrics can be deployed to assess the clusters' quality?
   (b) How can the segments generated be visualised and analysed?

(c) How could the model be improved in the future?

The remaining sections are organized as follows. In Chapter 2, a systematic literature review is presented, highlighting the main trends and recurring themes that were identified within the literature for both CS and EWS models. The third chapter, instead, discusses the methodology adopted to conduct the research and describes the guiding frameworks and analytical methods that were implemented to shape the study. Next, in Chapter 4, an overview of the experimental set-up designed to develop the customer clustering model is provided, whereas Chapter 5 introduces the results obtained for the clustering algorithm selected and explores the different segments generated. Finally, the limitations and future recommendations are presented in Chapter 6, which also outlines some final considerations and conclusions regarding the research outcomes as well.

2

LITERATURE REVIEW

2.1. METHODOLOGY

Scopus [1] was chosen as the main and only abstract and citation database not only for its voluminous collection of scientific articles, but also for the wide range of features it offers to its user to search for relevant literature [14].

The retrieval of the documentation was accomplished by making use of the advanced search bar, which allows researchers to insert a number of key words and create custom search queries. To download the research papers, it was decided to make use of the following key words: 'Early Warning Systems', 'Credit Risk', 'Financial Distress', 'Customer Segmentation', 'Customers Clustering', 'Unsupervised Machine Learning', 'Credit Portfolio Monitoring', 'Lending' and 'Loans'. The logical operators 'AND' and 'OR' were also implemented and combined with the above-mentioned search terms in order to create appropriate search queries. Since the research is focused on two main areas of interest, concerning the use EWS for credit risk monitoring and CS for credit portfolio management, three search queries were deployed to collect the literature regarding these topics and provide appropriate answers the research questions:

1. 'Early Warning Systems' AND 'Credit Risk' OR 'Financial Distress' OR 'Lending'
2. 'Customer Segmentation' OR 'Customers Clustering' OR 'Unsupervised Machine Learning' AND 'Credit Portfolio Monitoring' OR 'Lending' OR 'Loans'
3. 'Early Warning Systems' AND 'Customer Segmentation'

To download the initial batch of articles the search was applied on 'Keywords', 'Abstract' and 'Article Title'. The number of articles fetched as result of these queries is 266 research papers, 148 for the first query, 114 for the second query and 4 for the third one. Next, the analysis was restricted to only journals, reducing the number of articles to 134, 88 regarding EWS, 51 re-

[1] https://www.scopus.com

lated to CS and 2 concerning the combination of the two techniques. In the third phase, it was decided to focus only in the areas of Computer Science, Decision Sciences, Economics, Econometrics, Finance, Business, Management and Accounting. The selection was further extended by choosing the articles written in English and published between 2017 and 2023, collecting a total of 66 documents.
Finally, in the last phase, the remaining documents were screened by analysing the titles and abstracts in order to identify those that were deemed to be irrelevant for the scope of the project and exclude them from the final list.

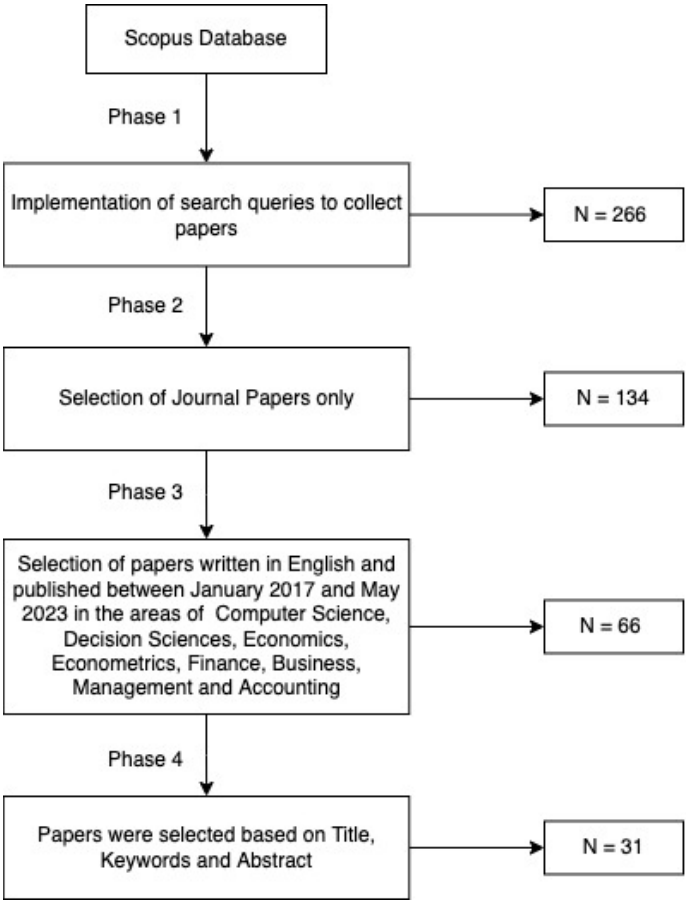


Figure 2.1: Articles selection process

The whole selection process is depicted in Figure 2.1, whereas the criteria that were implemented to achieve the final collection of articles are reported in Table 2.1.

2.2. RELEVANT TRENDS IN LITERATURE

This chapter contains three subsections: the first one will focus on the analysis of the year distribution of the published literature regarding the use of CS techniques in the lending field and the application of EWS for credit risk monitoring, the second subsection, instead, will investigate the journal distribution of the literature related to the two main topics. Finally, the third subsection will explore the most recurring and frequently-used key words.

Table 2.1: Summary of the criteria used to select the articles for the SLR

Criteria	**Decision**
The pre-defined key words are included in the title, abstract or in the key word list of the paper	Inclusion
The paper was published in a scientific journal	Inclusion
The paper was written in English	Inclusion
The paper was published before 2017	Exclusion
Duplicates of an original paper	Exclusion
The paper's abstract, title and content are not relevant to the research objective	Exclusion

All the articles that were consulted to conduct the current research are listed in Table B.1. As it can be observed, the table not only presents an overview of the papers collected, but also highlights several important aspects for each article, such as the settings in which the respective model was carried out, the main purpose of the experiment, the data-driven techniques applied and, finally, the evaluation methods used to determine the efficiency of the projects. The table was designed in order to enable the reader to fully understand how the findings of the current research were derived and facilitate the reading as well.

2.2.1. YEAR WISE DISTRIBUTION

Figure 2.2 presents the year distribution of the 13 articles related to the implementation of CS techniques in the lending field that have been thoroughly analysed to conduct the current research. The picture shows that, overall, the publications regarding this specific topic have been quite steady and stable throughout the last 6 years. However, the chart also indicates a notable increase of published literature in 2022, which may suggest a growing interest and use of clustering models among financial institutions. The observed growth is likely to be related and due to the technological advancements that this field of research has seen over the years, which led to more accurate and precise results compared to the previous attempts. Figure 2.3 portrays the same year distribution study for the literature concerning EWS for credit risk monitoring. As it can be observed, the literature published between 2017 and 2020 is quite limited and scarce. In fact, within the list of articles collected, there are no records of papers distributed respectively in the year 2018 and 2020.

On the other hand, similarly to the distribution of CS articles, it seems that from 2021 more and more researchers have started to acknowledge the potential that EWS hold and publish studies related to their implementation. 2022, in particular, represented the year with the highest number of publications. The reason behind the one paper from 2023 reported in this study, instead, may be due to the fact that the articles' selection for this systematic review occurred at beginning of the same year and, therefore, the literature available was still extremely narrow at that time.

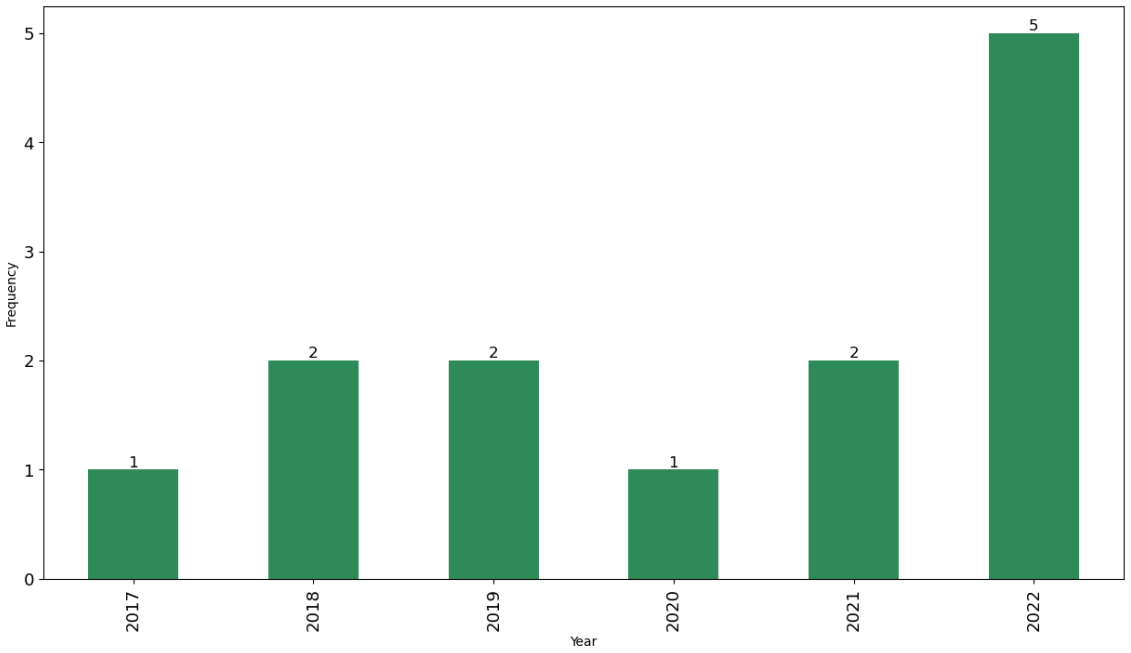


Figure 2.2: Year distribution of customer segmentation-related research papers

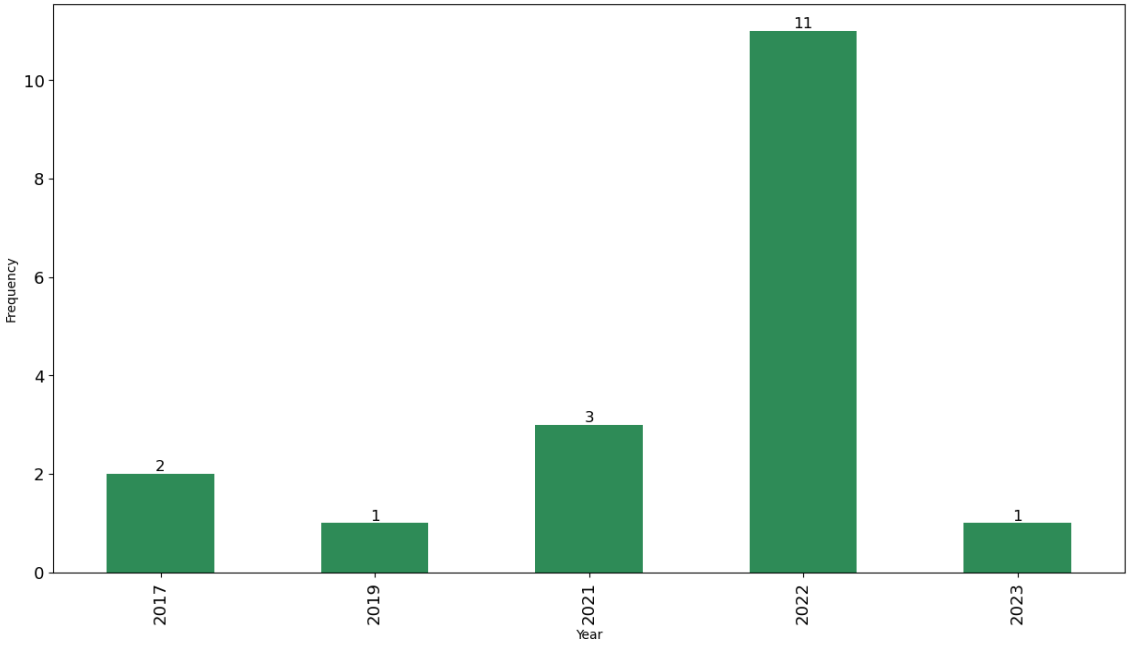


Figure 2.3: Year distribution of EWS-related research papers

2.2.2. JOURNAL WISE DISTRIBUTION

This next subsection aims at analysing the distribution of the literature collected across the different journals in which they were included.

Table 2.2 contains a list of all the CS-related articles retrieved, underlining the journal in which the article was published, the number of citations that each paper collected and the journal's

impact factor. A first observation that can be noted at a glance concerns the noticeable difference in the amount of citations among the different articles. In fact, it can be observed that two studies in particular present a significantly higher number of citations compared to the other articles [15, 16]. Moreover, the table also reveals that each paper was published in one singular journal and that no specific journal of preference could be identified for the researchers working in the field.
The “Impact Factor" column, instead, reports the Impact Factor (IF) of each journal concerning all the publications of the year 2021 and that, by definition, represents the journal’s most recent and updated IF score. In case the IF value was not publicly available at the time of the research, the information was omitted from the list.

Table 2.2: Summary of the journals distribution of CS-related articles

Study	Journal	Number of Citations	Impact Factor
Kaminskyi, A., Nehrey, M., Babenko, V., Zimon, G. [17]	Journal of Risk and Financial Management	–	2.3
Jadwal, P.K., Jain, S., Pathak, S., Agarwal, B. [18]	Microsystem Technologies	3	2.012
Machado, M.R., Karray, S. [19]	Electronic Commerce Research and Applications	–	5.622
Tasgetiren, N., Tigrak, U., Bozan, E., Gul, G., Demirci, E., Saribiyik, H., Aktas, M.S. [20]	Concurrency and Computation Practice and Experience	1	1.831
Yuan, K., Chi, G., Zhou, Y., Yin, H. [16]	Research in International Business and Finance	10	6.143
Pandey, K.K., Shukla, D. [21]	Reliability: Theory and Applications	–	0.44 –
Singh, In., Kumar, N., Srinivasa, K.G., Maini, S., Ahuja, U., Jain, S. [22]	Applied Soft Computing	5	8.263
Lazo, D., Calabrese, R., Bravo, C. [23]	Journal of Credit Risk	1	0.880
Morandi, S., Mokharab Rafiei, F. [15]	Financial Innovation	42	–
Nazari, A., Mehregan, M., Tehrani, R. [24]	International Journal of Supply Chain Management	–	–
Philip, D.J., Sudarsanam, N., Ravindram, B. [25]	Data Base for Advances in Information Systems	5	1.828
Firouzabadi, S.M.A.K., Taghavifard, M.T., Sajjadi, S.K., Soufi, J.B. [26]	International Journal of Electronic Customer Relationship Management	–	–
Luthfi, E.T., Wibowo, F.W. [27]	International Journal of Simulation: Systems, Science and Technology	–	–

For what concerns the EWS-related articles investigated, a similar trend to the one just examined emerges from table 2.3. In fact, it can be affirmed that the papers consulted belonged to multiple and distinct journals, with little overlap detected. However, it also seems that the "Mobile Information Systems" journal represents one of the most popular sources chosen by the experts, as it was associated with the largest number of articles.
In addition, also the distribution of the data related to both the number of citations appears to be almost identical to the previous case, though with slightly more distributed numbers and missing values.

Table 2.3: Summary of the journals distribution of EWS-related articles

Study	Journal	Number of Citations	Impact Factor
Wang, L., Zhang, W. [28]	Information Processing Management	–	7.466
Guerra, P., Castelli, M., Côrte-Real, N. [29]	Economic Analysis and Policy	4	4.444
Petropoulos, A., Siakoulis, V., Stavroulakis, E. [30]	Intelligent Systems in Accounting, Finance and Management	2	–
Wangsong, X. [31]	Mobile Information Systems	–	–
Xie, H., Shi, Y. [32]	Mobile Information Systems	–	–
Han, X. [33]	Computational Intelligence and Neuroscience	–	–
Yin, L.L., Qin, Y., Hou, Y., Zhao, J.R. [34]	Computational Intelligence and Neuroscience	2	–
Xie, W. [35]	Mobile Information Systems	–	–
Huang, B., Yao, X., Luo, Y., Li, J. [36]	Annals of Operations Research	4	4.820
Xu, L., Chen, W., Wang, S., Mohammed, B.S., Lakshmana Kumar, R. [37]	Annals of Operations Research	5	4.820
Tong, L., Tong, G. [38]	Scientific Programming	1	–
Yang, G. [39]	Computer-Aided Design and Applications	1	–
Jacobs, M. [40]	International Journal of Financial Studies	2	–
Zhu, L., Li, M., Metawa, N. [41]	Information Processing and Management	21	7.466
Aytaç Emin, A., Dalgıç, B., Azrak, T. [42]	Applied Economics Letters	1	1.287
Zhang, W., Chen, R.-S., Chen, Y.-C., Lu S-Y, Xiong, N., Chen, C.-M. [43]	IEEE Access	2	3.476
Pompella, M.,Dicanio, A. [44]	Economic Modelling	7	3.875
Berlinger, E. [45]	Finance Research Letters	–	9.846

2.2.3. KEY WORDS WISE DISTRIBUTION

This final subsection aims at providing an analysis of the most frequent key words that the authors have included in their articles. The purpose of key words is to facilitate the searching process of researchers and search engines in order to allow them to discover relevant articles in a more time-efficient manner. In general, key words define the field and subjects covered by the article and capture the most important aspect of the paper. However, it is important to acknowledge the fact that, due to the subjective nature of the choice of the key words used to collect articles, biases could be inherently introduced within the research. Still, in the scenario of this study, this phenomenon can be considered only partially significant and relevant since the application of the search queries did not only involve the papers' key words but also their respective title and abstract.

Nevertheless, it is important to underline that a remarkable number of scientific papers involved in the current research were not provided with key words at all, especially in the case of EWS. For this reason, the writers decided to exclude the articles in question from this analysis and consider only the ones that actually presented key words.

Figure 2.4 represents a word cloud of the most popular key words among CS-related papers that were examined during this study. Here, the size of the words is proportional to the frequency

of the respective word found in the literature. As it can be observed, the term “credit scoring" seems to be the most common key word used by the authors, followed by “k-means", “clustering" and “anfis". This discovery suggests that these topics have been addressed in more than one article and thus may represent recurring themes in this domain.

Figure 2.4: Word cloud of CS-related articles' key words

For what concerns EWS-related articles, the word cloud depicted in Figure 2.5 unveils a different scenario compared to the previous one. Here, the most relevant words that have appeared multiple times are, instead, “machine learning", “early warning systems", “financial crisis" and “deep learning".

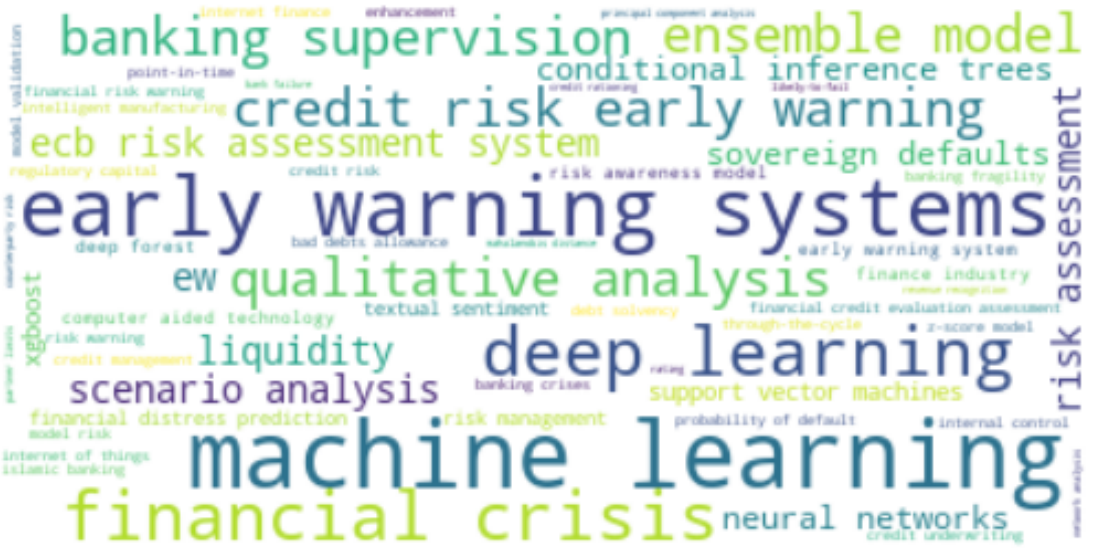


Figure 2.5: Word cloud of EWS-related articles' key words

2.3. DOMINANT THEMES IN LITERATURE

In this chapter, the dominant trends characterising the literature of EWS and CS are outlined. The analysis will be conducted from three perspectives involving the settings of the experiments, the techniques adopted and the evaluation methods applied.

2.3.1. SETTINGS ANALYSIS

This first subsection wants to provide an overview of the recurring settings that have been addressed in the literature related to both CS and EWS models. The term "settings" refers to the contexts in which the studies were conducted that strongly influenced the project's requirements and the subjects involved in the experiments.

With regards to the literature on CS, it was first discovered that a number of articles were focused on Peer-to-Peer (P2P) lending services, a form of financial technology that enables individuals to lend money or obtain loans from other individuals without the intervention of a financial institution [17–19, 21]. In fact, the Lending Club dataset, containing demographical and behavioural data belonging to the users of the Lending Club digital marketplace, was implemented in different researches to conduct and validate their experiment [18, 21]. However, it is important to underline that the majority of the articles that were reviewed were actually aimed at segmenting clients of banking institutions.

Another important aspect that emerged throughout the setting analysis, which also adds on to the points previously discussed, concerns the type of entities that have been taken in consideration for the investigation. It was observed that retail banking customers, consisting in individual consumers of the general public and population, represented the main subjects of the clustering models [15, 21, 24–27]. Indeed, the most frequent features and variables upon which the clusters were generated usually reflected the socio-economic nature of the subjects and, therefore, they were only applicable for the segmentation of real physical borrowers.

Finally, the analysis also highlighted the fact that almost all of the CS articles examined were deployed for risk assessment and monitoring purposes. Most of the paper reviewed, in fact, addressed to use of customer clustering techniques for the development of models aiming at identifying high-risk borrowers, characterised by a critical likelihood of dealing with financial distress. Only a limited number of papers implemented these techniques for different end-goals, such as better client-product allocation [26].

For what concerns the settings analysis of the literature on the use EWS for credit risk control, one important consideration must be outlined and discussed. Unlike what emerged during the investigation of CS techniques, the studies related to EWS targeted a wide range of industries and institutions. Although the field of Internet Finance and banks still represented a significant portion of the whole documentation [29, 31, 32, 42, 44, 45], most of the early warning applications analysed were developed in order to quantify and evaluate the financial risk of multiple types of corporates, including IoT companies, manufacturing enterprises and governments [28, 30, 33–35, 37, 41]. Moreover, as highlighted in Table A.1, a considerable amount of studies

were mainly focused on the Chinese market from a banking perspective and also in terms of general industries and firms [28, 31, 32, 36].

2.3.2. TECHNIQUES ANALYSIS

With the advancement of technology and the development of ever more innovative and state-of-the-art AI applications, nowadays, a vast number of solutions can be implemented to develop well-performing EWS and CS systems. Therefore, this subsection will examine the main techniques that have been adopted in the different studies in order to classify the most popular methods and obtain a better understanding of these systems from a technical aspect.
Regarding the literature related to CS applications, several considerations ought to be discussed. Firstly, it was discovered that CS relies on the application of one main technology: the use of unsupervised learning algorithms capable of identifying similarities within the entities and clustering borrowers that share similar characteristics from several perspective [15, 16, 18, 19, 25]. Among all the different clustering techniques addressed in the papers, two solutions have been deployed in multiple studies, and they are, respectively, K-means and Fuzzy C-Means (FCM) [15, 16, 19–21, 24, 26]. The first technology, in particular, featured almost all the clustering-related projects as its implementation is deemed to be extremely versatile, accessible and more efficient when benchmarked with other clustering algorithms. Nevertheless, since the performance of K-Means and other clustering approaches highly depends on the number of clusters selected, some authors made use of specific and renowned methods to guide the selection process. For instance, two projects made use of the Elbow Method to identify the appropriate clusters number [18, 19]. Other methodologies mentioned were the Ward Method or the Silhouette Score, which allow users to assess the clusters compactness and separation degree [26].
In addition, it was noted that a number of articles examined did not focus solely on clustering existing clients but, instead, they presented Hybrid Machine Learning (HML) models that originally aimed at realising different tasks [15, 16, 18, 19, 22, 23]. Within HML models, various types of data-driven techniques and algorithms work together in order to solve problems that they would not be able to solve independently [46]. For example, some clustering-related studies originally attempted to create models in charge of predicting clients' default risk. However, in order to further improve the accuracy of their artefact, clusters of similar borrowers would be initially generated so that they would act a starting point upon which the prediction would be executed.

For what concerns EWS-articles, the literature presents many facets regarding the type of applications developed and the techniques used. As Table A.1 shows, the concept of "Early Warning Systems" is not associated to one only and unique type of technology, but, instead, it involves multiple different practices. The most common implementation, which was discussed and deployed in several studies, consisted in the creation of a single risk indicator predicting the default risk of customers. Within this context, two different methodologies have been used in particular. The main one, adopted in the majority of the articles, involved the benchmarking of

numerous supervised Machine Learning algorithms so that the most accurate and performing model could be chosen [28–30, 36, 39]. The second one, instead, involved the use of statistical measurements and formulas [41, 42, 45].
Regarding the first methodology mentioned, a variety of algorithms have been employed within the literature, including Logistic Regression, Support Vector Machines, Random Forest and Decision Trees. Yet, the algorithm that proved to be the most successful and efficient is XGBoost, an implementation of Gradient Boosting Decision Trees, praised for its executional speed and accuracy of the results [29, 30, 36, 39].
On another note, several other EWS-related studies have delved into the development of more comprehensive EWS that went beyond the sole prediction of credit risk. Some of the articles consulted, in fact, aimed at presenting articulated credit risk management architectures that were comprised of a number of different modules and processes, each one executing a specific function and task [31, 32, 35]. Other articles, instead, aimed at introducing new extensions to existing EWS, such as additional risk indicators and indexes, in order to improve their risk management power [42, 45]. Because of this variety of different EWS applications, it can be observed that the field of EWS appears to be an emerging and developing domain, still lacking of one unique and common framework.

2.3.3. EVALUATION METHODS ANALYSIS

The purpose of this final subsection is to provide an analysis of the dominant and trending evaluation methodologies that researchers have adopted in order to validate the performance of their models and to assess their efficacy.
For the articles concerning solely customer clustering models, one validation metric in particular appeared in two different studies and was used to determine the quality of the clusters built, the Silhouette Score. In general experts tend to rely on more than one indicator to assess the performance of their clustering algorithm, however the Silhouette Score was the only common index that was exploited in few researches [21, 24].
Regarding the literature related to hybrid models, instead, it is important to remind the reader that the applications involved did not propose the clustering of customers as their initial objective. Therefore, during the evaluation processes, the authors were actually interested in examining the artefact's performance with regards to the experiment's final outcome. For this reason, the quality validation of the clusters obtained as an intermediate step of the overall procedure was often omitted or not explicitly discussed in these papers [15, 16, 18, 19].

As for EWS-related articles, the evaluation methods employed depended mainly on the type of application developed. In the previous subsection, it was discussed that a substantial part of the literature in this field proposed models capable of predicting the default risk of subjects on the basis of labeled data. Consequently, the methodologies used to evaluate these systems consisted in the calculation of various metrics and in the exploitation of a few validation techniques commonly adopted in classification-based projects. The most recurring metrics detected were

the AUC Score, Accuracy Score and the F1 Score, however several other metrics have appeared in various studies with less frequency, such as the Kolmogorov Smirnov Score or the G-Mean Score [28, 30, 36]. Moreover, researchers have also relied on two major validation techniques to evaluate the performance of the predictors, involving train/test split and k-fold cross validation. In the latter case, in particular, the value of the variable k was usually selected at the author's discretion based on the dataset characteristics and computational resources.
In the case of the articles concerning more sophisticated and complex architectures functioning as actual EWS, no specific and common indicators were found within the literature. It was observed, however, that the assessment of the effectiveness of these structures, in some occasions, was based on qualitative analyses of the infrastructures and the potential impacts that these new systems would have on existing processes [31, 43].

2.4. THE RELATIONSHIP BETWEEN EWS AND CUSTOMER SEGMENTATION SYSTEMS

Given that one of the research questions of this systematic review aimed at discovering potential links in the literature concerning CS projects and EWS, this section will try identify these connections and discuss them in detail. Before delving into the exploration, it is first important to remark that none of the papers investigated openly addressed or tackled this subject in explicit manner. Indeed, even when researching for articles inherent to both EWS and CS, no relevant results could be obtained from the Scopus' collection.
Nevertheless, in the subsection 2.3, describing the most widespread and used techniques, it was underlined that many of the CS-related studies that have been consulted consisted in the development of certain hybrid Machine Learning models, combining clustering and supervised learning algorithms to make predictions related to customer's default risk. At the same time, it was also discussed that on several occasions the term "EWS" was associated with models also aiming at determining the default risk of the subjects. The analysis, therefore, suggests the existence of a subtle relationship between these two technologies. In fact, it can be deduced that clustering algorithms can play a relevant role in the creation of EWS, as they can be implemented within default prediction models by detecting patterns and risk profiles within the data and by integrating this information into the financial risk estimation process. On the other hand, as it can be observed from Table A.1, there are no instances of articles reporting the integration of early warning indicators within CS models, entailing that this sort of application has not been documented or deployed yet.
In conclusion, following the observations provided in this analysis, it can be affirmed that, despite the fact that no study has focused on uncovering the link between these two techniques, a relevant connection can be indirectly derived and elaborated. However, the connection in question does not involve the application of early warning signals into CS systems, which seems to represent still an unexplored domain.

3

METHODOLOGY

This chapter embarks on a comprehensive analysis of the methodology employed in the current CS study. Methodology serves as the guiding framework that shapes the entire research process, ensuring rigor, clarity, and effectiveness in achieving the study's objectives. The methodology chapter is divided into three distinct sections: the Design Science Research, the Cross-Industry Standard Process for Data Mining and the Analytical Methods sections. The former two sections delve into the exploration of the Design Science Research (DSR) and CRoss Industry Standard Process for Data Mining (CRISP-DM) methodologies, providing in-depth descriptions of their principles and applications in the context of this study. The Analytical Methods section, instead, focuses on the technical aspects of the study's methodology. It outlines the data-driven techniques and models utilized to implement the customer segmentation application. These methods are rooted in leveraging the power of data to uncover meaningful patterns and insights within the customer dataset.
By dividing the methodology chapter into three separate subsections, the study aims to provide a complete overview of the whole research framework.

3.1. DESIGN SCIENCE RESEARCH

DSR is a systematic and innovative approach tailored to address complex problems in the field of Information Systems (IS). It encompasses a number of principles and practices that are deemed to be essential when conducting a research in the context of IS [1]. The term "Design Science" refers to the process of designing or investigating a state-of-the-art artifact that would improve the environment in which it is integrated and generating design knowledge to obtain technological advanced solutions for real-life problems [47]. Therefore, since the aim of the current study is to create a new customer segmentation artifact to identify potential business opportunities using early warning signals, the DSR method appears to be the most appropriate and suitable for this research.

An overview of the whole DSR methodology process model is depicted in Figure 3.1.

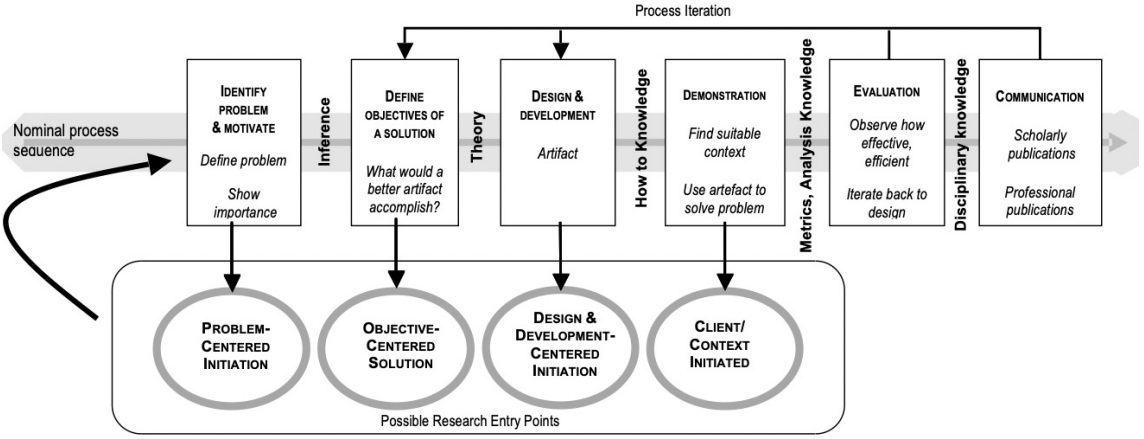


Figure 3.1: DSR Methodology Process Model (Source: Peffers et al. [1])

As it can be observed, the process is divided into six main activities [48]:

1. **Problem identification**: The background motivations and main research problem must be identified and addressed in order to justify the value of the solution that is being proposed and allow the audience to better comprehend and accept the results of the research.

2. **Objectives definition**: The objectives and goals inferred from the problem specification have to be properly outlined and defined. They can be either quantitative or qualitative.

3. **Design and Development**: This step requires the determination of the desired functionalities and architecture that the artifact should present and, subsequently, the actual development of the solution designed.

4. **Demonstration**: Once the artifact is designed and developed, the efficacy of its implementation must be demonstrated. This can be achieved either by actually deploying the model in the real-life scenario or by studying its application in case studies and simulations.

5. **Evaluation**: In order to verify whether the research goals are indeed achieved, the results obtained from the deployment of the model must be compared with the initial objectives defined during Step 2. At the end of this activity, the researcher should be able to understand if a re-visitation of the model created is necessary to further improve its efficacy or if the “Communication" step can occur.

6. **Communication**: Finally, all the key concepts emerging from the implementation each step of the DSR Methodology must be properly communicated to the main stakeholders involved with the respective project, when relevant. The most appropriate forms of

communication depend on the context of the project and the audience involved, however, scholarly and professional publications represent the most common means used in research.

Since customer segmentation involves handling with large and diverse datasets and uncovering hidden meaningful patterns, the approach proposed in DSR is well-suited to deal with the complexity that this domain presents and producing solutions with real-world applications. Moreover, the adoption of the DSR methodology for this specific project ensures that the process is systematic and rooted both in terms of research and practical design principles, which would guarantee more valuable and accurate outcomes for the organisation.

3.2. CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING

In addition to the Design Science Research methodology, the study also followed the CRISP-DM process model, developed in 1996 in order to tackle the complexity and guide the development of data mining-related projects [49]. A "data mining" process is defined as the automated procedure of identifying patterns within a large set of data and, from these findings, extracting business insights and predictions [50]. The CRISP-DM standard, outlining the most appropriate transformations and steps that should be implemented throughout the respective project, was established in order to support any data mining task.
Similarly to what was presented for the DSR methodology, an overview of the whole process model of the CRISP-DM methodology is depicted in Figure 3.2.

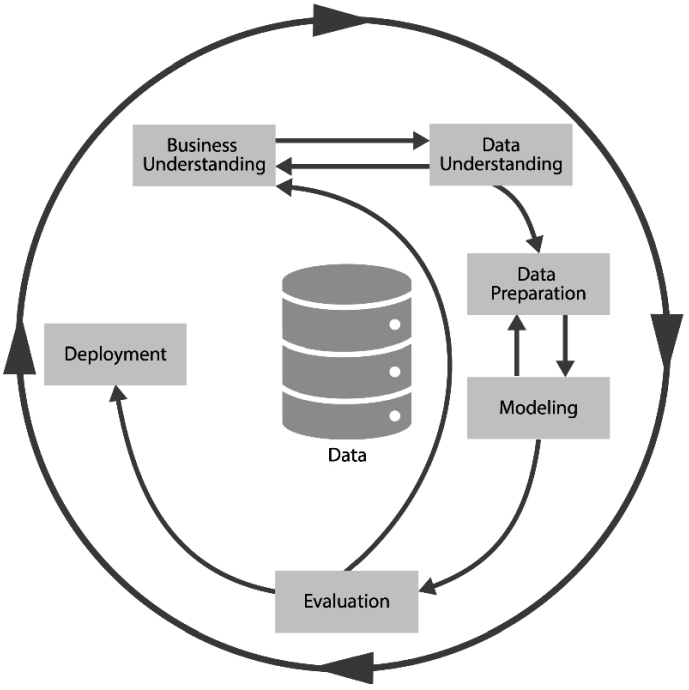


Figure 3.2: CRISP-DM Process Model (Source: Kristoffersen et al. [2])

Once again, the model is divided into a number of phases that represent the life cycle of each

data mining project. Despite the chronological order of the steps, portrayed by the arrows indicating the inter-dependencies among the different phases, the sequence is not to be deemed rigid and compulsory in real-life scenarios. Indeed, moving back and forth between the multiple steps is often required and necessary [2]. In addition, the outer circle depicted in the figure is supposed to symbolise the cyclical nature of each data mining process.
A brief description and analysis of each passage of the methodology is provided in the following subsection [2, 49]:

1. **Business understanding**: As an initial step, all the relevant project stakeholders, including business analysts and data scientists, must collaborate in order to define the project's objectives and requirements from a business perspective. This knowledge is, then, converted into a proper definition of the business problem and a preliminary plan.

2. **Data understanding**: Next, the project team is in charge of collecting and exploring the data in order to identify potential quality issues and become familiar with the completeness of the dataset.

3. **Data preparation**: During the data preparation step, all the activities aimed at constructing the final dataset from the initial raw data are performed. Some of the activities in question can involve data cleaning, transformation, and feature engineering before the implementation of the modeling tools.

4. **Modeling**: Once the "Data preparation" phase is complete and the final dataset is obtained, the different analytical models selected are deployed and calibrated to achieve their optimal performance.

5. **Evaluation**: All the models built in the previous phase must evaluated and reviewed with the purpose of determining whether the project's results are, indeed, aligned with the initial business objectives defined in the "Business understanding" phase. At the end of this specific step, the decision on the model to deploy as a final data mining result should be reached.

6. **Deployment**: The final phase of the CRISP-DM methodology involves the "go live" of the final model in the respective production environment. However, it is important to underline that this last step highly depends on the business requirements that have been established in the previous phases, involving, for instance, simple business reports or complex data-driven infrastructures requiring repeatable monitoring.

In conclusion, within the context of the application of EWS for the automated segmentation of WB customers, the CRISP-DM methodology provides a structured and effective approach that guarantees an appropriate preparation of the data required and a proper development of the most suitable clustering model. By following this framework, the authors were capable of ensuring that the final artifact presented all the necessary characteristics and contributions with regards to the initial business preliminary conditions and goals.

3.3. ANALYTICAL METHODS

3.3.1. PRINCIPAL COMPONENT ANALYSIS

Dimensionality reduction is crucial in many data analysis and ML tasks. High-dimensional datasets often suffer from the curse of dimensionality, where the increased number of features can lead to increased complexity and inaccurate results [51]. Principal Component Analysis (PCA) plays a significant role in addressing these challenges. PCA can be described as a statistical method that reduces the number of variables of a table to its so-called Principal Components (PCs), representing its essential features [52]. This technique consists in projecting the data into a new orthogonal system of coordinates that retains the most significant information and patterns while also reducing the dimensionality of the dataset and the impact of random variations and outliers. The goal of PCA is to be able to explain the maximum amount of variance with the fewest number of PCs, which are ranked in order of importance [52]. Indeed, the first reported principal component captures the most significant variation in the data, the second captures the second most significant variation orthogonal to the first, and so on.
Some other advantages that can be obtained from the deployment of PCA are also [51]:

- **Reduced complexity and higher computational efficiency**: Since PCA retains only the most important information while reducing dimensionality, it can lead to faster computation and improved models implementation.
- **Elimination of multi-collinearity**: As previously mentioned, the principal are orthogonal to each other, meaning they are uncorrelated. Therefore, this property enables the elimination of multi-collinearity which can result in unstable results.
- **Better data visualisation**: PCA's ability to reduce data to two or three dimensions makes it suitable for clusters visualisation. By plotting data points in the reduced space, complex relationships and clusters can be more easily observed and understood.

Nevertheless, like any method, PCA presents also number of limitations and drawbacks that users need to be familiar with in case they consider implementing this technique. Among the several disadvantages, the most significant and relevant ones are the following [53, 54]:

- **Linearity and correlation assumptions**: The algorithm highly relies on the relationship between the variables of the dataset. Specifically, PCA assumes that the features in input present a linear inter-correlation. For this reason, the method is not well suited for capturing non-linear relationship and determining the main PCs of non-correlated variables.
- **Sensitivity to scale**: PCA is defined as "scale variant". This means that the algorithm tends to be biased towards the features that report bigger values and wider standard deviations compared to the other variables. A common solution to this limitation involves the standardisation of the original dataset to a common and unit standard deviation, in order to avoid the domination of a single variable over the others.

- **Loss of interpretability**: As the main PCs represet a linear combination of the features characterising the original dataset, the actual interpretation of the resutls obtained from the implementation of the algorithm can be rather challenging and trivial.

3.3.2. K-MEANS

In the realm of customer segmentation, understanding the diverse characteristics and behaviors of customers is crucial to tailor marketing strategies and optimise customer experiences. K-Means, a powerful and widely used clustering algorithm, plays a pivotal role in achieving this goal. It is described as one of the simplest and most common clustering algorithms, praised for its ability to deal with large amount of data with efficient computational time and accurate results [55]. K-Means is an iterative unsupervised clustering algorithm that partitions the data into K pre-defined and non-overlapping clusters, where each data point belongs to one and only cluster. The algorithm is expected to reach the local optimal and guarantee the highest similarity and intra-cluster closeness [55]. In order to achieve this final result, K-Means implements a number of steps [56]:

1. Initialisation of K random data points as the initial clusters centers;
2. Calculation of the distance of each data point from every cluster center using the sum of squared errors;
3. Assigning each data point to its nearest cluster center;
4. Calculation of K new centroids by computing the average of all the data points that belong to each; cluster

Step 2, Step 3 and Step 4 are repeated iteratively until convergence, meaning that the cluster centroids do not move any longer. Nevertheless, as a localised optimisation method, the outcome of the K-Means clustering algorithm is sensitive to the starting centroids of the initial clusters and the number of segments selected [55]. Therefore, several are methods proposed by the literature for choosing the correct K. One of the most renowned approaches used is, in fact, the Elbow Method. The main goal of this method is to select the point of diminishing returns, i.e. the elbow point, when visualising the Within Cluster Sum of Squares (WCSS), the average sum of squared distance between each data point and its cluster's centroid, for different numbers of clusters. The intuition behind this approach is that clusters will naturally improve in the fit and in compactness as the number of clusters increases [57]. However, at some point, in proximity of the so-called elbow point, the increase will be overfitting the model and no significant changes will be visible in terms of WCSS.

3.3.3. DBSCAN

Although K-Means represents one of the most efficient and performant clustering algorithms, no single algorithm is best for all possible purposes. Since the performance of a clustering algorithm highly depends on the contextual circumstances and the type of data available for

segmentation, it is important to test and benchmark different models in order to identify the most appropriate one for the project's end goal. Therefore, in the current research, the results obtained from the deployment of K-Means have been compared with the clusters generated from the application of an alternative clustering method, DBSCAN.
Several aspects and characteristics differentiate the two algorithms and their implementation. Firstly, DBSCAN does not require the user to provide a fixed number of clusters. Instead, as a density-based clustering application, it determines the number of clusters based on the density and neighbourhood relationships of the data. Secondly, since all K-Means clusters present spherical and convex shapes, K-Means may fail to identify non-linear relations. On the contrary, DBSCAN is capable of forming clusters of arbitrary shapes which may be more well-suited for the type of dataset used. Finally, DBSCAN can efficiently manage and ignore noise by treating these data points as individual points, part of clusters of a lower density [58].

Nevertheless, the implementation of DBSCAN relies on the definition of two important parameters: ϵ and MinPoints. The former represents the length of a radius that is used to determine the neighbour data points of each data point. The latter, instead, defines the minimal number of points that are necessary in order for a point to be considered a "core point". Data points that contain less than MinPoints or no data point within their radius are called, respectively, "border" and "noise" data points [59, 60]. Once all the core, border and noise points are labeled, the algorithm is in charge checking each data point one by one. If the data point is, indeed, a core point, then a new cluster is formed containing all points within its ϵ distance. Points that are reachable from these points are included in the cluster as well. If the point is not a core point but is within ϵ distance from a core point, it is classified as a border point and associated with the cluster of the nearest core point. Instead, if the point is labeled as a noise data point, the algorithm will considered it to be an outlier [60].

3.3.4. SHAP VALUES

Nowadays, despite the fact that ML models are becoming more and more performant and computationally efficient, the algorithms behind these applications have also faced an increase of their complexity, which lead to a significant depreciation of their explainability and understandability. Therefore, today, such algorithms tend to be considered as black-boxes [61]. Nevertheless, interpreting the predictions made by complex ML models is crucial for building trust, diagnosing issues, and extracting insights, especially in the case of sensitive domains, such as healthcare or finance, where the decisions derived from these models can lead to severe consequences. To overcome this difficulty, the SHAP (SHapley Additive exPlanations) method provides a innovative framework for explaining the output of these models. The method, which is based on cooperative game theory, represent a unified approach to compute the sum of the contributions of all the individual features on the model's final prediction [62]. Some of the key concepts of this methodology are the following [61]:

- **SHAP Values**: SHAP values attribute to each feature and quantify the actual change that would be visible on the model's prediction when conditioning on that particular feature.

According to Shapley [63], the SHAP value can be defined as a method that assigns a certain payout to each player based on the contribution that the player has on the total payout.

- **Local Explainability**: SHAP values explain a feature contribution on the model's prediction for one specific instance.
- **Global Explainability**: The aggregation of the SHAP values obtained for multiple different instances can provide insights on the features that have consistent impact on the models behaviour and performance.

Several methods can be used to combine the insights and approximate SHAP values. Two approximation methods, respectively the Kernel SHAP and Shapley sampling values, are model-agnostic. However, four other model-type-specific approximation methods have been integrated in the approach as well, i.e. the Linear SHAP, the Low-Order SHAP, the MAX SHAP and the Deep SHAP [61].
Python offers a new powerful and widely used library, the SHAP library, capable of automating the calculation of SHAP values across various ML models, including tree-based models, linear models, and more. The application of the SHAP library is not limited to supervised machine learning models. It can also be employed to make unsupervised models more interpretable and understandable. While traditional supervised SHAP implementations assign values to individual features, in the case of unsupervised clustering algorithms, the SHAP values calculated aim to define the contributions of features to the overall clustering algorithm procedure and functioning.

4

EXPERIMENTAL SET-UP

4.1. EXPERIMENTAL SET-UP

The aim of this research is to identify potential business opportunities within the WB lending credit portfolio by implementing a CS model that generates clusters of similar clients based on historical records of the early warning triggers raised and entities' internal data. With regards to this initial research objective and the background motivations that lead to the development of this project, several requirements have been derived, defining the main characteristics that the model should present:

- **Risk-oriented**: since ARIA is capable of generating early warning triggers and anticipate potential risk events, the use of the information collected from the application to implement the CS model will consequently lead to the clustering of entities that share similar risk levels and the development of risk-oriented segments. Financial and behavioural-related features are not included in the current study, however, future research could investigate on the integration of these aspects within the model to obtain even more accurate and insightful clusters.
- **Identifiable**: it is important that specific groups of entities, with distinct characteristics and properties, can be recognised and identified within the clusters generated.
- **Actionable**: in order to ensure that the model can actually serve as a tool that facilitates managers' decision-making and opportunities identification processes, actionable insights should be derived from the analysis and exploration of the segments created.

The process depicted in Figure 4.1 provides a graphical overview of all the mechanisms that have been deployed in order to achieve the final results. The phases concerning the non-technical procedures, such as the definition of the business objectives and requirements, which are not addressed in the overview, were still part of the project's framework and were reported

in the previous chapters and sections.
As it can be observed, the investigation was divided into three main processes involving the collection of triggers and clients data, which were already part of the initial framework of the EWS, the data preprocessing and the analysis of the clustering results obtained, which, instead, represented the innovations and experiments introduced by the current research. Each process was, then, composed by a number of more specific and sequential sub-processes that were implemented to realise the main procedure. All the processes and sub-processes above-mentioned were developed through the use of a number of Python libraries and tools supported by Jupyter Notebook[1], a web-based development environment for creating data science projects. In addition, from a more in-depth exploration of the schema, it can be noticed that the three main processes represented the central phases defined in both the DSR and CRISP-DM methodologies. More specifically, the "Data Collection" and "Data Preprocessing" processes related to the model's "Design and development", and "Demonstration" steps of the DSR method, and the "Data Ubderstanding", "Data Preparation" and "Modeling" steps of the CRISP-DM method. The phase concerning the clusters analysis and examination, instead, can be regarded as the "Evaluation" step that both process models include.
Every procedure highlighted in the overview is thoroughly described and analysed in the following sections, outlining also the motivations behind certain critical decisions that have deeply influenced the outcome of this research.

[1] https://jupyter.org

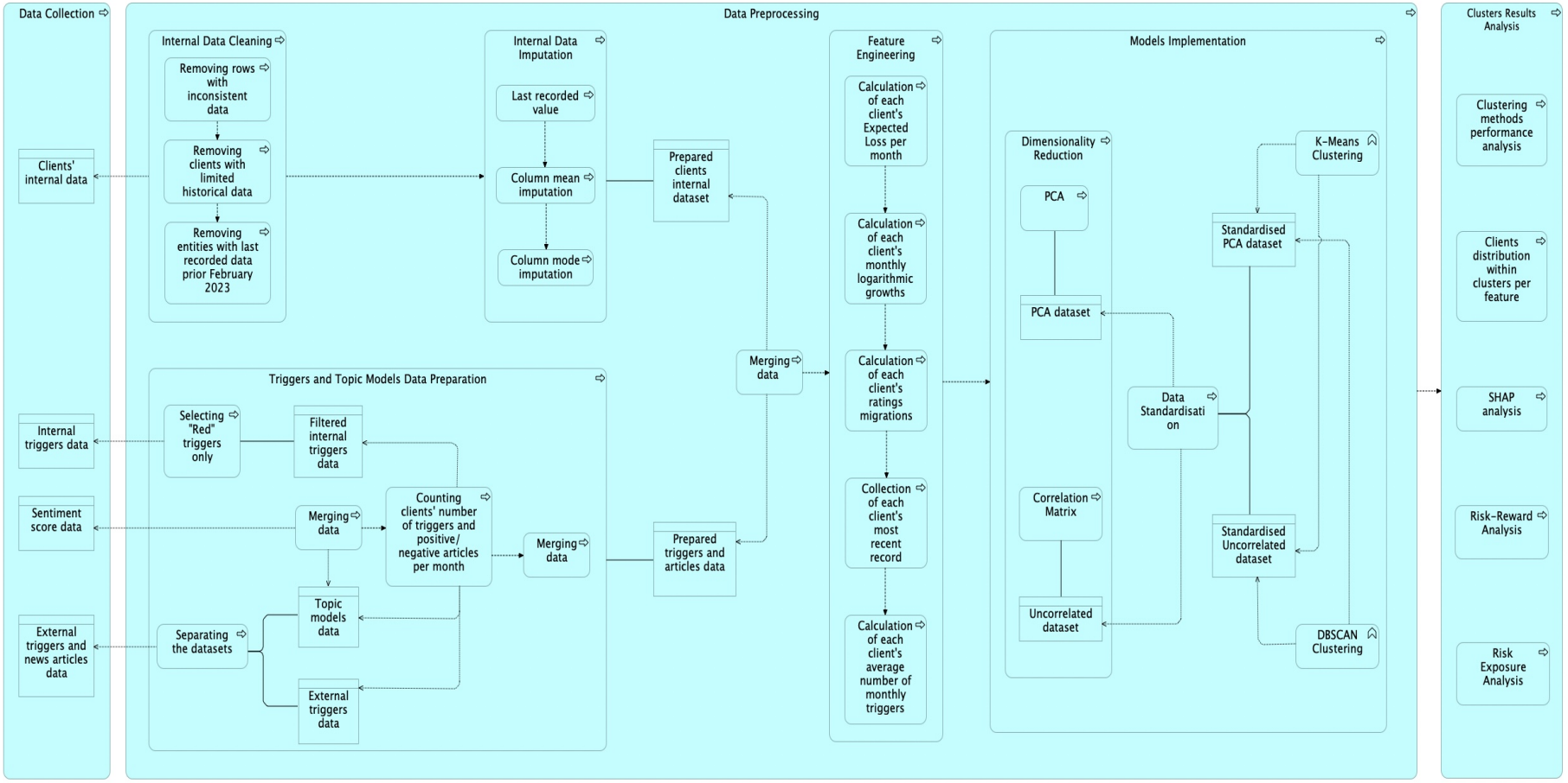


Figure 4.1: Overview of the Experimental Set-Up

4.2. DATA COLLECTION

The data used to develop the project was provided by ING and concerns the same type of the data processed by the ARIA team to monitor customers financial health and generate the early warning triggers. As already mentioned in the first chapter, the ARIA team has developed three different pipelines to extract and manipulate the data in question. Firstly, an internal pipeline was built in order to collect clients' monthly data from ING central financial and risk systems. The information obtained is, then, calculated to produce the so-called "internal triggers". Secondly, two other separate external pipelines were developed with different purposes. One pipeline is in charge of retrieving public market data related to specific macroeconomic indicators and financial products from Refinitiv [11]. The second pipeline, instead, extracts news articles regarding multiple relevant topics from online newspaper or news aggregator services, such as the Financial Times [64] and Google News [65]. These two infrastructures enable the calculation and generation of the "external triggers".

Therefore, three datasets have been retrieved from ARIA's databases in the form of CSV files, all containing historical records of triggers and internal data collected from May 2022 to April 2023. The 12-month period of time in question was selected in order to achieve a yearly overview of customers' risk profile changes and migrations. However, based on businesses' emerging necessities and objectives, different time windows could be implemented in future studies.

A complete description of each table is presented in the next three sub-sections, however, due to confidentiality reasons, only general qualitative information of the clients will be provided. Moreover, the insights obtained from the exploratory analysis of the data retrieved and additional relevant details, such as data descriptive statistics, are thoroughly discussed in Chapter 5.

4.2.1. EXTERNAL TRIGGERS

The external triggers table includes records of all the external triggers that the ARIA application was able to raise from May 2022 until April 2023. This dataset not only contains triggers activated by the detection of significant changes for certain financial instruments, such as Equity or Credit Default Swap (CDS), but also reports of all the news articles involving the monitored entities discovered by the EWS. A secondary table was also included in the study, containing the respective sentiment scores that each articles was associated with. With such measure, quantifying the tone and nature of the articles in question, stakeholders can obtain timely and quantitative knowledge of how the organisations are perceived by the public and make more informed decisions on the actions to implement. In order to facilitate the understanding of the concept of external triggers and news articles, also known as topic models, Table 4.1 provides an overview of all the external triggers managed by ARIA and brief descriptions of how they are raised.

Table 4.1: Summary of all the external triggers included in the study

Trigger Type	ID	Description
External Trigger	BND	Raised when the Bond value changes over a pre-defined threshold
External Trigger	CDS	Raised when the CDS value changes over a pre-defined threshold
External Trigger	EQU	Raised when the Equity value changes over a pre-defined threshold
External Trigger	FXR	Raised when the Foreign Exchange Rate EUR/VAL value changes over a pre-defined threshold
External Trigger	ECR	Raised when the External Credit Rating value changes over a pre-defined threshold
Topic Model	BNK	Flagged when a relevant Bankruptcy-related news article is detected
Topic Model	ECC	Flagged when a relevant Environment & Climate Change-related news article is detected
Topic Model	FRD	Flagged when a relevant Fraud news article is detected
Topic Model	HR	Flagged when a relevant Human Rights-related news article is detected
Topic Model	MA	Flagged when a relevant Merger & Acquisition-related news article is detected
Topic Model	SNC	Flagged when a relevant Sanctions-related news article is detected

4.2.2. INTERNAL TRIGGERS

Similarly to the previous dataset, the second table processed contains historical data of the internal triggers flagged by the early warning application over the same period of time considered in the external triggers table. As already mentioned in the initial section, these types of triggers are derived from specific financial measures that ING internally calculates for each client. Thanks to the internal pipeline, the information is then retrieved from ING's central systems and stored into ARIA's. Again, Table 4.2 provides an overview of all the internal triggers involved in the study.

4.2.3. INTERNAL DATA

The final table used to conduct the experiment is the internal data table, consisting in monthly records of client's personal information of both demographical and financial nature. The original dataset presented a significant number of features, some of which were deemed to be not in line with the objective of the research. Therefore, all the columns that had been labeled as irrelevant were discarded from the study and are not reported in Table 4.3.

Table 4.2: Summary of all the internal triggers included in the study

Trigger Type	ID	Description
Internal Trigger	FBS	Trigger generated when a significant change in the Forbearance Status is detected
Internal Trigger	ESRT	Trigger generated when a significant change in the ESR Transaction Outcome is detected
Internal Trigger	SS	Trigger generated when a significant change in the Sanction Status is detected
Internal Trigger	CVNT	Trigger related the Covenant Monitoring and scheduling
Internal Trigger	DPD2	Trigger generated when the Days Past Due value changes over a pre-defined threshold
Internal Trigger	IR/RG2	Trigger generated when the Internal Rating Grade value changes over a pre-defined threshold
Internal Trigger	EAD2	Trigger generated when the Exposure At Default value changes over a pre-defined threshold
Internal Trigger	LE2	Trigger generated when the Outstanding Amount value exceeds a pre-defined threshold
Internal Trigger	RWA2	Trigger generated when the Risk-Weighted Assets value changes over a pre-defined threshold
Internal Trigger	LGD2	Trigger generated when the Loss Given Default value changes over a pre-defined threshold
Internal Trigger	ROD/RUD	Trigger generated when the Reviews Upcoming Date is after/before the respective Reporting Date
Internal Trigger	IFRSS2	Trigger generated when a change in the International Financial Reporting Stage (IFRS) status is detected
Internal Trigger	WL	Trigger generated when the difference between the current Reporting Date and a Watchlist Date is below a certain threshold

4.3. DATA PREPROCESSING

In any data-driven study, the processes of data preprocessing and feature engineering play a pivotal role in ensuring the accuracy and reliability of the subsequent analyses. In the context of this very research, these procedures represented the most time-consuming and meticulous phases of the project. The complexity lied in the multifaceted nature of the data collected, which reflected different aspects of the customers and their risk profiles. Given also the significant number of incorrect instances and outliers within the dataset, the crafting and refinement of the features to extract and capture all the various desired nuances not only required a certain level of domain expertise, but also general technical capabilities and knowledge. This chapter delves into the fundamental steps taken to transform the raw data into an optimized dataset, laying the groundwork for the future segmentation investigation.

The data used to develop this study was derived from an yearly collection of historical records, encompassing ARIA's triggers and internally calculated client data, carefully collected throughout the period spanning 2022 and 2023. This dataset formed the bedrock of the current analy-

Table 4.3: Summary of the internal data included in the study

Trigger Type	ID	Description
Internal Data	ALLOC LIMIT AMT	Client's total allocated limit
Internal Data	OUTSTANDING AMT	Client's total outstanding amount
Internal Data	EAD	Client's Exposure At Default value
Internal Data	RWA	Client's Risk-Weighted Assets
Internal Data	WORST IFRS	Client's worst credit risk stage based on IFRS9 accounting standards
Internal Data	LGD	Client's Loss Given Default value
Internal Data	PD	Client's Probability of Default value
Internal Data	ACTIVE STATUS	Client's current Activity Status
Internal Data	RISK RATING	Client's current Internal Risk Rating

sis, allowing the author to draw valuable insights and make informed decisions. As discussed in the previous section, the data was sourced from three distinct datasets: internal triggers data, external triggers data, and client-specific data. Each of these datasets was initially deployed independently and individual preprocessing steps were applied to cleanse and enhance their quality. Once these preprocessing steps were completed, the datasets were merged into one unified final dataset.

The following subsections will meticulously expound upon each preprocessing steps undertaken, offering an in-depth understanding of the operations deemed necessary to prepare the data for clustering purposes. Since transparency and comprehensibility are essential elements in building confidence and clarity around the research's findings, the report will provide clear explanations regarding the rationale behind each transformation.

4.3.1. TRIGGERS AND TOPIC MODELS DATA PREPARATION

To prepare the external, internal triggers and news articles, also known as topic models, for further analysis, several rather simple and elementary preprocessing procedures were undertaken.

Firstly, the initial challenge with the external triggers dataset related to the fact that the table contained information on both external triggers and external news articles detected for specific entities. To address this issue, the first preprocessing step involved the separation of news articles from the external triggers data. This separation was achieved by filtering the data based on the reported trigger ID. By isolating the news articles from the rest of the data, a distinct table, specifically dedicated to the topic models, was obtained. As for the internal triggers, instead, the data in question was also filtered by selecting the "Red-flagged" triggers only. This triggers colour schema designed by ING has been implemented in order to define the magnitude of risk involved when a specific data change occurs. Therefore, by selecting the "Red" internal triggers only, it was decided to consider only the signals representing the most critical and potentially dangerous events.

Next, the following step of the preprocessing procedure focused on counting the number of occurrences of all the triggers and topic models flagged for each customer on a monthly basis. This counting process was approached differently for the triggers and the news articles. For the external and internal triggers tables, the data was grouped based on the customer's ID, record month, and record year. This grouping allowed the determination of the number of warnings raised for each specific entity within a given period of time. By aggregating the data in this manner, valuable insights into the frequency and intensity of triggers reported for each customer throughout the dataset's timeline were gained. The handling of the topic models involved a more nuanced approach instead. First, thanks to the integration of an additional dataset provided by the team containing sentiment scores related to various articles, the sentiment of each article were associated with the corresponding news article. With the inclusion of this information, it was possible to assign either a positive or a negative sentiment to each news article in the topic models table, based on the respective sentiment score. Finally, similarly to what has been done to the triggers tables, the data for regarding topic models was grouped based on the client's ID, record month, and record year. This grouping allowed the calculation of the number of positive and negative articles detected for each client and every month of activity.

4.3.2. INTERNAL DATA CLEANING

Client's internal data table posed more unique and complicated challenges during the preprocessing phase of the research. In fact, the dataset required a more extensive data cleaning procedure compared to the previous tables, as it contained a number of incorrect and empty rows that demanded thorough examination and cleansing. The following steps outline the data cleaning process undertaken to ensure the integrity of the dataset:

1. **Identifying and Removing Rows with Inconsistent Data**: The initial step involved analysing the dataset to identify records containing inconsistent data. Rows would be considered incorrect if they met two specific criteria. In particular, a row was labeled as incorrect if the reported Outstanding Amount was equivalent to zero but the corresponding EAD or Risk-Weighted Assets (RWA) were greater than zero. Likewise, a row was flagged as incorrect if the Outstanding Amount was greater than zero, but the respective EAD or RWA for that month were equal to zero. These inconsistencies could be due to data entry errors or system issues and, therefore, were deemed unsuitable for accurate analysis.

2. **Removing Entities with Limited Historical Data**: In the context of this study, the goal was to segment entities on the basis of their progress and activity over the months. For this reason, the entities with only one single record in the dataset were considered to have too limited historical data. Since meaningful analysis of their growth and progress would be hindered by this lack of data, it was decided to exclude such entities from the clustering process. This step ensured that the segmentation analysis focused on clients with sufficient historical information to derive valuable insights.

3. **Eliminating Clients with Most Recent Record Preceding February 2023**: As a final step

of the cleaning process, clients for whom the most recent record in the dataset preceded February 2023 were excluded. The motivation behind this exclusion was based on the experts' suggestion that the absence of data for the last three months of the study would indicate that these entities had terminated their loan contracts with ING and no longer had any open loan to repay. Since the primary objective of the study was to identify potential up-selling opportunities within the existing customer base, it was deemed inappropriate to include clients who were no longer part of the credit portfolio at the time of the study.

By implementing these rigorous data cleaning steps, the internal dataset was refined and prepared for subsequent analysis. The removal of inconsistent and irrelevant data ensured that the remaining records accurately represented the relevant clients' information. This meticulous cleaning process was essential in providing a reliable foundation for the clustering and segmentation analyses, ultimately contributing to the generation of meaningful and actionable insights for the study.

4.3.3. INTERNAL DATA IMPUTATION

Following the data cleaning process, a systematic data imputation process was employed in order to address the missing values in clients' internal dataset. The imputation aimed to replace the missing values with reasonable estimates to maintain the dataset's integrity.
First, by grouping rows related to the same customer ID, both forward and backward filling techniques were applied. This way, all the missing values within a customer's data were filled with the last or first available non-null value. This approach ensured that the imputed values were consistent with the chronological order of the records for each customer. For the remaining missing values in numerical columns, such as the Outstanding Amount, EAD or RWA features, the imputation was performed by replacing the NaN cells, indicating an undefined or an unrepresentable value, with the overall mean of the respective column. Calculating the overall mean was necessary in order to approximate the missing values with a measure that reflected the central tendency of the numerical distribution. Finally, in the case of missing values for the one and only categorical column of this study, i.e. the Active Status feature, the imputation was carried out using the overall mode for that column. The mode represents the most frequently occurring value in a categorical column, therefore, the implementation of this technique maintained the distribution of categories within the dataset.

As most clustering algorithms are not capable of managing missing information [66], this data imputation process represented an essential step in producing a comprehensive and reliable dataset. By employing these imputation techniques,it was ensured that the clients' internal dataset was be complete and suitable for clustering analysis.

4.3.4. FEATURE ENGINEERING

The Feature Engineering step is a crucial part in the data preprocessing pipeline of every Machine Learning-related research. It consists in transforming the data available into new vari-

ables that are more suitable for modeling and analysis [67]. This process is of a paramount importance for every customer clustering study as not only it allows to integrate domain business knowledge to generate variables that are aligned with the segmentation's initial objective, but it can capture important features and patterns that are not directly visible as well.

After the data imputation stage, the data was still distributed across four distinct tables, all sharing a common format. Each row was indexed based on customer ID, record year and record month. To proceed with the feature engineering step, the data stored in these tables had to be merged together based on the above-mentioned indexes. An outer join was initially applied between the triggers and articles tables, followed by a left-join with the client's internal dataset. This merging process produced the final comprehensive dataset suitable for subsequent feature engineering.
In this study, the feature engineering process was divided into a number of minor sub-processes which were aimed at implementing different manipulations to specific types of columns individually. Since the data contained in the final dataset represented time-series data in the form of monthly records, it was important to leverage each client's historical information so that the knowledge regarding the entity's progress and evolution over time could be obtained:

- **Creation of the "Expected Loss" variable**: Initially, a new feature was created, the Expected Loss variable. The Expected Loss can be defined as a credit risk parameter derived from the multiplication of the EAD, PD and Loss Given Default (LGD). The introduction of this new metric aimed at reducing the dimensionality of the dataset and aggregating the information of three individual variables, which would not be as interpretable if included separately, to enhance the quality of the insights that would be derived from the segments.

- **Computation of the Monthly Growths Using Logarithmich Models**: For the columns concerning the EAD, RWA, Outstanding Amount, Allocated Limit and Expected Loss the growth between the entity's starting date and ending date was computed by calculating the difference of the logarithmic values for each attribute. In order to handle cases of undefined or infinity growths caused by the specific nature of the logarithm function, the values recorded at starting date and ending date of each client that were reportedly equivalent to 0 were set 0.01. Moreover, to ensure a fair comparability across clients with varying starting dates, the growth was divided by the number of months of activity of the respective client. The use of the logarithms to calculate growth rates ensured that, in case of large growth and percentage changes, the use of log units can provide more symmetric and accurate representations of the changes [68].

- **Computation of Credit Risk Status Migrations**: The focus of this sub-process was on the Internal Risk Rating and Worst IFRS Stage columns. Firstly, for both features, a new threshold was set, defining good and bad levels of credit, based on pre-defined criteria established by the two standards. Clients with scores below or equal to this threshold were considered to have reasonable and potentially profiting credit quality, while scores above

the threshold indicated increased credit risk. The threshold for the Worst IFRS Status column was equivalent to Stage 1 of the standard. As for the Internal Risk Rating standard, the information regarding the set threshold can not be openly shared for confidentiality reasons.

On the basis of these two threshold, four new variables for both features were created, representing all the possible status migrations: positive no migration (from good to good status), negative no migration (from bad to bad status), positive migration (from bad to good status), and negative migration (from good to bad status). This process allowed the tracking of the credit quality changes that have occurred over time for each entity. In particular, if a client, in its first month and last month of activity, was reported to have, respectively, a “bad" and a “good" score, then the positive migration column for that specific client would indicate the value 1, whereas the remaining migration columns would be assigned with the value 0.

- **Collection the Most Recent Record**: Since the identification of potential up-selling and business making prospects is usually not solely based on the progress and improvement that a client has made, but also on the actual status of the client's current loan, the most recent records were extracted for the columns representing the Outstanding Amount, Total Allocated Limit and Activity Status. The data retrieved corresponded to the values reported in the last month of activity of each client.

  Moreover, the Activity Status column was renamed to “Default Watchlist Status" as it was transformed into a binary column indicating whether the client had been labeled as “In Default" or “Watchlist".

- **Computation of Average Number of Monthly Triggers and News Articles**: Finally, as a last Feature Engineering step, the average number of monthly internal triggers, external triggers, positive articles, and negative articles was computed for every entity and corporate.

4.4. MODELS IMPLEMENTATION

This subsection delves into the implementation of the clustering algorithms, which are instrumental in segmenting customers and extracting valuable insights from the dataset. Before diving into the specifics of the clustering algorithms, two essential preparatory steps were applied to the final dataset obtained: dimensionality reduction and data scaling. With the application of these two techniques, the effects of dimensionality and distribution of the data on the accuracy of the clusters are mitigated, leading to more insightful and coherent segments.

For the current study, two widely-used clustering algorithms have been adopted, K-Means and DBSCAN. Each algorithm offers distinct advantages and is well-suited to different types of data and clustering objectives. In the forthcoming subsections, all the in-depth details and motivations behind the decision to use these two clustering models will be provided and outlined.

4.4.1. DIMENSIONALITY REDUCTION

Since a high number of input features can often make the clients segmentation difficult to interpret and challenging, one of the last and final steps before the actual implementation of the clustering models involved the reduction of the dimensionality of the dataset. Fewer input dimensions correspond to a fewer number of input parameters that could also lead to simple clustering models [69]. This was achieved by adopting two different methodologies. The first method involved manual feature selection based on the results of variables' correlation analysis. The correlation analysis aims to assess the relationship between two features by computing a decimal value called the Pearson's correlation coefficient, calculated using the following formula:

$$C_{A,B} = \frac{Covariance(A,B)}{\sigma_A \sigma_b}$$

If the coefficient is greater than zero, then the two variables in question have a positive correlation, meaning that an increase of one variable would also lead to an increase of the second variable [70]. This analysis was carried out for every tuple of features included in the dataset by plotting a correlation matrix, which would show how strong the relationship between each tuple of features is.
Features that exhibit high correlation with one another may contain redundant information which can hinder the interpretation of the clusters generated. Therefore, by identifying and removing both positively and negatively highly correlated features, only the most diverse and informative attributes in the dataset can be retained. For this reason, all the attributes that presented important correlations and that were deemed to be negligible for the research objective were excluded.

Alternatively, another widely-used dimensionality reduction technique in customer segmentation projects is PCA. As already mentioned in the previous chapter, this method aims to identify a set of orthogonal axes, known as principal components, capturing the directions along which the data exhibits the most significant variation. The projection of the data onto these components reduces its dimensionality while preserving the data's essential patterns and initial structure. In this study, the strategy employed to determine the optimal count of principal components focused on the exploration of the percentage of variance that would have been retained by varying numbers of principal components. This involved generating a graph depicting how much of the variance is accounted for by each component, as well as how the combinations of the different components add up to the total variance. Consequently, only the components capable of explaining between 80% and 95% of the overall variance were taken in consideration.

4.4.2. FEATURE SCALING: DATA STANDARDISATION

Feature Scaling is an essential preprocessing step in Machine Learning, especially when applying distance-based algorithms such as clustering. Since real-life datasets usually contain data of varying magnitude, variance and ranges, it is important to make sure that all the features are

on a comparable scale in order for the ML model to interpret them correctly. Feature scaling, in fact, prevents certain features from prevailing and lead to a biased model by transforming all variables to a similar scale. Several techniques can be used to achieve this result, including Min-Max Normalisation and Standardisation. However, since the data collected presents a significant number of outliers and noise which could impact the original distribution of the data, it was decided to make use of the Standardisation technique, as it is also capable of maintaining the relationships between data points [71].
Standardisation, also known as Z-Score Normalisation, is a scaling method that transforms the data so that mean of the attribute becomes zero and the resulting distribution has a unit standard variance [71]. The formula deployed is the following:

$$X' = \frac{X - \mu}{\sigma}$$

Where μ represents the mean of the feature and σ the respective standard deviation.

4.4.3. K-MEANS

As previously addressed in Chapter 3, after preprocessing and standardising the final dataset, the first step required for implementing K-Means clustering was to determine the optimal number of clusters. In this study, the Elbow Method guided this decision-making process. The technique in question involves running K-Means for a range of cluster numbers and plotting the corresponding WCSS values against the number of clusters. In general, a higher number of clusters usually leads to a lower WCSS value, as each point is closer to its centroid. However, adding too many clusters can hinder the interpretation of the segments generated. The term “Elbow" derives from the shape of the line depicting the WCSS values against the number of clusters. Initially, the WCSS value decreases steeply as the number of clusters rises. The rate of decrease starts to slow down after the so-called elbow point is reached, determining the point of inflection on the curve and representing the potential candidate for the optimal number of clusters.
Although this method proves to be a valuable approach for selecting the most appropriate K, several challenges can be encountered when implementing it. In fact, it can occur that the elbow point might not be easily discernible in the plot in case of absence of a distinct point of inflection. In order to address this issue, the “KElbowVisualizer" tool, offered by the Python library Yellowbrick, was incorporated in the implementation of K-Means. The tool can accurately establish the elbow point in an automated manner, even when its identification might be visually difficult to perceive.

The performance of K-Means clustering , however, depends also on the definition of a number of other hyperparameters, which play a critical role in shaping the outcomes of the algorithm. In addition to the number of clusters, the remaining hyperparameters that have been experimented in the study are [72]:

- **Initialization Method**: The “init" hyperparameter defines how the initial cluster centroids are initialized before the algorithm starts its iteration. The value assigned to this hyperparameter was “k-means++", an initialisation method that selects the initial centroids “based on an empirical probability distribution of the points' contribution to the overall inertia", leading to faster convergence and better results compared to random initialisation.[72].
- **Number of Initialisations**: The “n init" hyperparameter specifies the number of times the K-Means algorithm will be run with different initial centroid seeds. The value assigned, in this case, was set to the default value of 10.
- **Random Seed**: The “random state" hyperparameter defines the random number for the initialisation of the centroid. Setting a specific integer ensured that the randomness was deterministic and the same results could be obtained when running the algorithm multiple times. The number chosen was 42, as it represented one the most common and popular integers.

4.4.4. DBSCAN

In contrast to K-Means, where hyperparameters like the number of clusters could be determined using methods such as the Elbow Method, selecting optimal values for DBSCAN's key hyperparameters, namely ϵ and MinPoints, presented more challenges.
Various studies have adopted different methods for selecting ϵ and MinPoints. The choice behind these hyperparameters were either derived based on the authors' domain knowledge or by adopting specific rule of thumbs. For example, a recurring approach involved setting the MinPoints hyperparameter equal to a predefined K value and plotting the K-Nearest Neighbor distances of each data point in ascending order to identify the so-called “knee" point that, similarly to the elbow point in the Elbow Method, corresponded to the optimal ϵ [73]. Unfortunately, such methods can be inadequate, particularly when applied to specific datasets with unique characteristics. In fact, in the context of this research, these methods proved to be rather inefficient and unsuitable when implemented. Therefore, given the limitations imposed by these traditional techniques, a more subjective approach to selecting ϵ and MinPoints was adopted. The focus of this analysis shifted toward finding a balance between the generation a reasonable number of clusters that captured meaningful customer segments and the optimisation of the respective Silhouette Score, a metric used to evaluate the quality of clustering results. This was achieved by plotting two matrices that provided insights into the performance of different parameters combinations. Both plots presented a grid of values for different ϵ and MinPoints combinations. In the first matrix, each cell in the matrix displayed the number of clusters formed by applying DBSCAN with the corresponding ϵ and MinPoints. The second matrix, instead, depicted the Silhouette Score for each parameter combination. By plotting these matrices side by side, it became possible to observe patterns and trends across different parameter combinations. Moreover, a visual inspection allowed for the identification of regions where a balanced trade-off between the number of clusters and the Silhouette Score was

guaranteed.

4.5. MODELS VALIDATION

In this final section, all the validation procedures adopted to provide a comprehensive exploration and evaluation of the customer segments generated are outlined. The analysis was developed from two main perspectives. Firstly, the quality of the clusters is assessed by investigating the cohesion and separation of the segments. This is achieved by calculating three popular measurements the help determine the best partition of entities. The second study, instead, thoroughly examines the characteristics of the customer segments based on all the features that have been maintained in the datatset. Descriptive statistics, such as mean values and distributions of features within clusters, are believed to provide meaningful insights into the distinct behaviors and growths of each segment. Finally, by integrating and interpreting the information obtained from the two previous analyses, the research aims to align the clusters with domain knowledge and business objectives in order to identify a group of entities that would represent ideal prospects for future business opportunities.

4.5.1. CLUSTERS QUALITY EVALUATION

To estimate the quality of the customer segments generated by the K-Means and DBSCAN clustering algorithms, three key metrics were employed: the Silhouette Score, the Calinski-Harabasz Score, and the Davies-Bouldin Score. These measures are defined as following [74, 75]:

- **Silhouette Score**: The Silhouette Score measures how close each data point is to its closest cluster. It is calculated using the mean intra-cluster distance, indicating the closeness of the points in the same cluster, and the mean nearest-cluster distance, defining the separation of points belonging to different clusters, for each sample. The measure ranges between the score -1, representing incorrect clustering, and $+1$ for highly dense clusters.
- **Calinski-Harabasz Score**: Also known as the Variance Ratio Criterion, is defined as “the ratio between the within-cluster dispersion and the between-cluster dispersion". A high C-H score entails that the observations within each cluster are close together, meanwhile clusters themselves are actually far away from each other. Therefore, the greater the score is, the better the performance.
- **Davies-Bouldin Score**: The Davies-Bouldin Index can be described as the average similarity between clusters, where the similarity is measured from the comparison between the size of the clusters and the within-cluster distance. As it can be discerned from this definition, a low D-B score signifies a more marked separation among the clusters.

As previously mentioned, the use of these three indicators provided valuable insights on the cohesion and separation of the segments created. Moreover, these metrics were computed for both the uncorrelated dataset, obtained from manual feature selection, and the PCA-generated dataset, in order to verify which dimensionality reduction technique was the most appropriate

for the study objective.

4.5.2. SEGMENTS EXPLORATION

The second type of analysis, instead, wanted to focus on the exploration of the clusters that have been developed in order to discover the main characteristics of each customer segment from several perspectives.
Initially, the analysis of the number of entities falling into each customer cluster helped identifying segments with higher density. Clusters with higher density signified a larger concentration of customers sharing similar characteristics, which may have also represented the most influential group of clients. Moreover, the study of the number of clients composing each cluster provided relevant insights on the actual performance of the clustering algorithm. For instance, if the clusters distribution of the entities proved to be extremely unbalanced, with certain segments containing very few clients, then it was discerned that the clustering algorithm adopted was not particularly efficient and failed to properly detect patterns within the data.
The use of descriptive statistics for all the features used in the clustering process, instead, shed light on to the main tendencies and spread of the data within each segment. Mean, median, standard deviation, minimum, and maximum values allowed the identification of the features and traits that distinguished a group of entities from the others and highlighted the distinct attributes that characterised the respective segment. Additionally, the investigation of the outliers and noise data points informed the researchers on the discrepancies and irregularities found in every clusters, which led to a deeper understanding of the effectiveness of the models implemented. However, since the presence of these instances may have also been caused by potential data quality issues or inappropriate preprocessing steps, researchers were also able to re-design their artifact in order to develop more accurate and uniform clusters.

Nevertheless, the exploration of the segments generated extended beyond the mere analysis of the descriptive statistics and entities distribution. Indeed, in order to shed light on to performance of the clustering algorithms and provide stakeholder with a holistic perspective on the clusters' effectiveness, a risk-reward analysis and an risk exposure analysis were also integrated in the study. The former investigation exploited the power of the average number of monthly triggers as a pivotal metric. The risk element was computed by aggregating the monthly averages of internal triggers, external triggers, and negative articles identified within each cluster. On the contrary, the reward factor was established by calculating the average count of monthly positive articles associated with each cluster. The use of these metrics enabled the identification of clusters that were more susceptible to financial volatility and unfavorable sentiments, while, at the same time, reflecting on the favourable aspects and positive market sentiments related to each segment.
The second investigation, instead, delved into a deeper assessment of the risk profiles of the generated clusters by examining the contrast between the average monthly Outstanding Amount growth and the average monthly EAD growth characterising every cluster. Since a more substantial escalation of the observed average EAD growth could signify an elevation in the asso-

ciated risks, the juxtaposition of these metrics allowed the researchers to quantify the genuine risk exposure inherent within each cluster. Moreover, where possible, the results obtained were compared with the Total Outstanding Amount reported in April 2023 in order to provide a more in-depth understanding of potential vulnerabilities that could arise when dealing with a certain cluster of entities.
It is important to stress the fact that, when calculating the above-mentioned averages, the outliers, which could have skewed the results obtained, were excluded from the calculation by taking in consideration only the entities falling within the interquartile range for each specific feature. This judicious approach aimed at ensuring integrity and objectivity in the data used.

The findings of these customer clusters explorations served as a foundation for strategic decision-making, empowering the bank to develop extensive customer-centric initiatives and stay competitive in a dynamic market landscape. Through this analysis, the study was able to reveal which high-value customers exhibited favorable attributes, such as a decrease of default risk and positive status migrations, and who were also likely to qualify for premium loan products or credit limit increases. Moreover, the insights collected also provided information on the entities that, at the time of the study, were facing repayment difficulties and financial distress hardships, which could serve as tool for ING to offer more targeted financial counseling and adjust their lending terms.

4.5.3. SHAP ANALYSIS

The SHAP library was harnessed in order to unravel the web of feature contribution that define the clusters generation process. This tool is capable of providing a better understanding of how the individual features influence the assignment of each entity to a particular segment. The computed SHAP values, in fact, represent a quantitative measure that captures the feature's impact on the clustering model. In the current study, the visualisation of SHAP values was executed independently for each cluster produced by K-Means, considering the dataset obtained from the dimensionality reduction process using correlation analysis. Although the SHAP library encompasses a number of different explainers that are specifically tailored to the type of supervised ML algorithm implemented, it poses a challenge for unsupervised ML models due to a lack of direct applicability. To overcome this limitation, the model-agnostic Kernel SHAP explainer was adopted in this research. This model has been developed to operate in an effective manner regardless of the type of ML model used in the application. A secondary limitation that emerged throughout the experimentation of the library related to demanding computational resource and complexity associated with fitting the explainer on the entire labeled dataset. To address this issue and optimise the efficiency of the model without compromising the integrity of the dataset, an alternative solution was devised. Instead of attempting to compute SHAP values for the complete dataset in a single run, the researchers adopted a more thoughtful and time-efficient strategy. The approach in question consisted in running the model multiple times on a smaller subset of the initial dataset, containing approximately 1,000 random samples each time. By breaking down the computation into smaller subsets of

the data, a more manageable utilisation of the SHAP explainer was achieved, while preserving the essence of the analysis. Then, by comparing the values obtained for the same cluster label across different runs, the consistency and robustness of the SHAP values generated was verified. This iterative process helped ensure that the final results produced did not represent random fluctuations, but, instead, were based on solid foundations and evaluations, ultimately confirming the reliability of the findings.

5

Results and Discussion

5.1. Exploratory Data Analysis

Before delving into the full analysis of the results obtained from the implementation of each analytical model, it was crucial to initially investigate the underlying structure of the data related to the features generated from the feature engineering process, outlined in Chapter 4. This procedure, referred to as Exploratory Data Analysis (EDA), serves the purpose of extracting information regarding the primary patterns and traits within the dataset, which allow to make better informed decisions about the best clustering approach to use. Two distinct analyses were implemented to explore the dataset from different perspectives, as depicted in Figure 5.1, Figure 5.2 and Figure 5.3. The former analysis (Figure 5.1 and Figure 5.3), explored clients distribution across different ranges of values of each feature, by quantifying the number of observation and instances discovered within various intervals of the specific attribute. However, given the large dispersion that some attributes presented and the imbalance in terms of the number of occurrences identified within each interval, comprehending the data distribution solely through the study of the histogram visualisations posed important challenges. Indeed, by inspecting the bars reported of Figure 5.3, it can be immediately noticed that, for specific variables such as the "tot_outstanding" or "negative_articles" attributes, only one singular column is observable. This does not entail that, for the respective features, one unique value was detected. Instead, it wants to highlight the fact that the majority of the clients fell within the reported range of values shown by the individual bar depicted. Still, a limited number of instances were characterised by unusual values that did not belong to the standard interval. Because this subgroup of clients represented only a restricted minority compared to the overall client base, the respective columns and bars are, therefore, not directly visible in the chart. To support the interpretation of the data, an additional analysis, involving the feature's descriptive statistics and reported in Figure 5.2, was implemented. Based on the illustrations provided in the aforementioned figures, the following insights and conclusions are drawn:

- ***Default/Watchlist Status***: As of April 2023, less than 2000 entities out of all the 22331 borrowers included in the study were associated with either a "In Default" or "Watchlist" activity status. Therefore, as expected, this subset of entities, composed by individuals that were incapable of fulfilling their financial obligations or that were characterised a number of risk factors that could potentially lead to default, represented only a minority of the overall number of clients.

- ***Worst IFRS stage and Internal Rating score migrations***: For what concerned the clients distribution for all the possible Internal Rating score migrations and Worst IFRS stage migration, it appeared that, for both standards, the majority of the clients that were included maintained a favourable credit quality and low credit risk by reporting an Internal Rating score and IFSR stage below the pre-defined thresholds from the beginning. On the contrary, the borrowers characterised by elevated ratings and either a Stage 2 or Stage 3 of the IFRS standard formed the second most prevalent group. In addition, the histograms indicated that a larger portion of the overall client base was subjected to a downward migration in their personal rating and assigned stage rather than an upward one. This suggested that clients were generally more prone to face a deterioration of their credit quality and financial health and less likely to achieve significant improvements.

- ***Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths***: Concerning the aspect related to monthly growth patterns, the descriptive statistics values associated to the different features revealed various relevant scenarios. Firstly, both figures suggested that, for a small subset of the entities involved, extreme decreases and increases were recorded across all features. Indeed, the customers in question presented values above 1 and below -1, representing growth rates exceeding, respectively, 100% and -100%. The underlying reason behind these unusual results can be attributed to the approximation approach that was implemented during the feature engineering stage to handle cases where the entities presented null values at the start or end of the period. Because of this data manipulation, the logarithmic growths computed for these specific subgroup of customers were particularly prominent.
  Despite this phenomenon, it was observed that, in reality, the majority of the borrowers experienced growths within the range of approximately -0.5 and $+0.5$, although some features presented narrower intervals as well. For instance, it appeared that, for the variables related to the Expected Loss growth and the Allocated Limit growth, most client reported increases that fell within the range of -0.2 and 0.2. Nevertheless, Figure 5.2, indicated that, in reality, for the Allocated Limit, the EAD and the Outstanding Amount growth features the larger portion of clients either faced minor declines over the months, especially for the former two attributes, or showed no substantial growth. This conclusion was drawn from the examination the features' median and 75^{th} percentile values, all of which were approximately 0 for these attributes.
  In addition, regarding the Expected Loss and RWA growths, it was discovered that, although a relevant number of clients faced positive growth in these features, the growths

for most of these entities were not notably significant. In fact, for both features, the 75^{th} percentiles depicted in Figure 5.2 were visibly negligible and did not indicate critical increases.

- ***Average number of monthly triggers and news articles***: At first glance, it was observed that the biggest portion of the entities involved in this study would not generate any external trigger, negative article or positive article warnings every month. This was, once again, derived by the analysis of the 75^{th} percentile value of these specific features (Figure 5.2), which, in all three cases, was equivalent to 0. Still, a minimal subset of clients was reported to raise between 1 or 2 triggers and articles on a monthly basis.
  For what concerned the internal triggers, the situation depicted in Figure 5.3 appeared to be slightly more concerning. Although most entities presented an average number of monthly internal triggers below 0.57, hinting that they had a minor likelihood of activating any internal trigger every month, over 5000 borrowers did report a greater and more significant average, which, for a substantial number of clients would even involved the activation of more than 1 trigger each month.

- ***Total Outstanding Amount and Total Allocated Limit***: Unfortunately, because of the significant deviation of the data related to both features, from the investigation of Figure 5.3 no meaningful insights concerning the clients' data distribution could be derived. As a result, the analysis predominantly relied on the findings presented in Figure 5.2.
  Firstly, it was noticed that, as of April 2023, approximately 50% of the clients presented a Total Outstanding amount below 5000€. Within this subset of entities, approximately 25% of the clients involved were characterised by a null value, suggesting that the customers had fully settled their outstanding balance or didn't utilise their allocated limit in the first place. Conversely, the remaining clients were characterised by larger balances, some even reaching values as high as 7×10^{-10} for specific individuals.
  Similarly, in the case of Total Allocated Limit, it was discovered that 50% of the borrowers presented less than 500000€, and, among them, 50% were characterised by limits below 5000€. The maximum value reported for this specific attribute was equal to 1×10^{-11}, which was most likely assigned to those entities that presented the largest outstanding amounts.

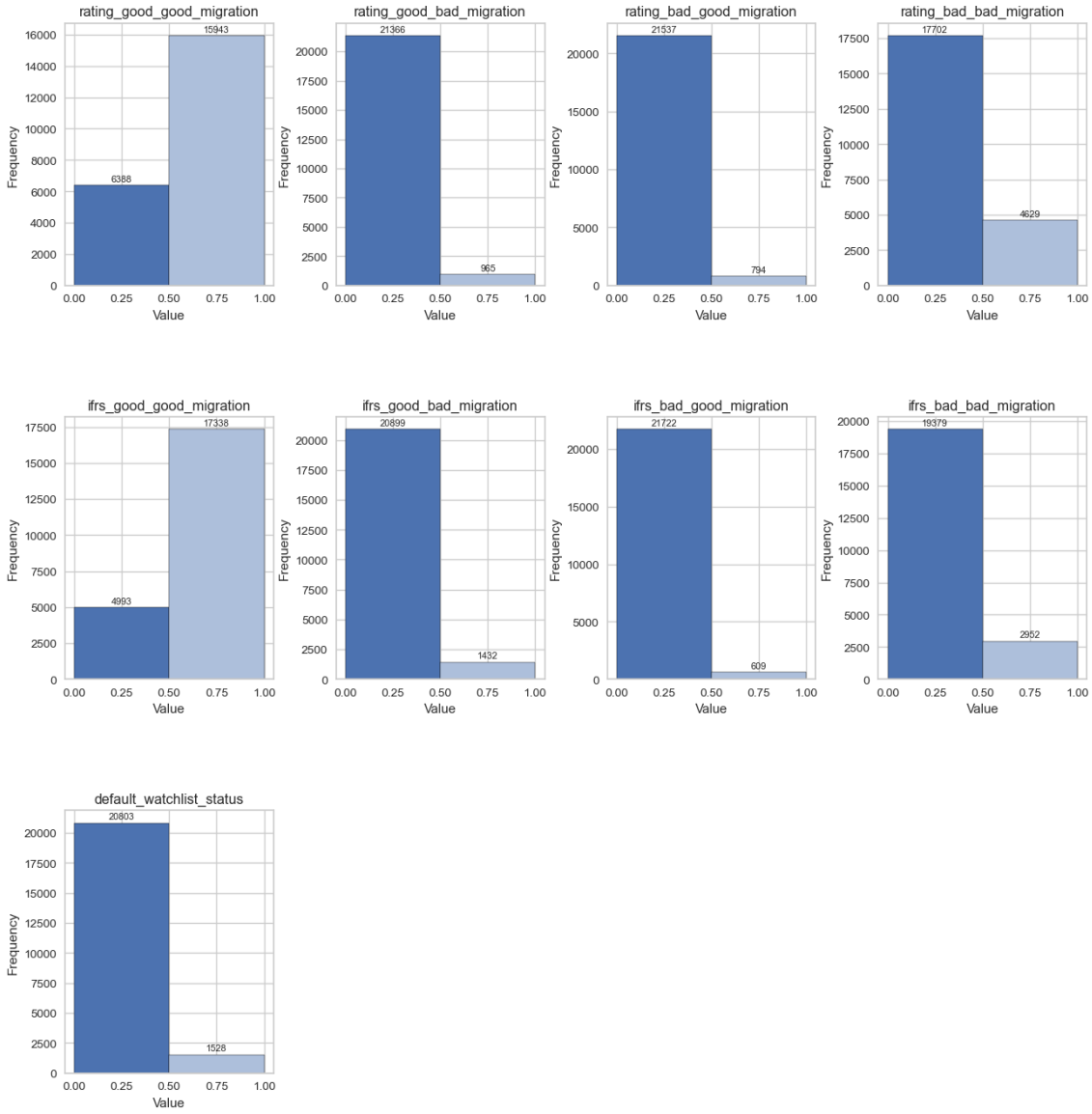


Figure 5.1: Clients distribution for different values of the migrations and Default/Watchlist status

	count	mean	std	min	25%	50%	75%	max
alloc_limit_growth	22331.0	-9.124462e-03	4.182453e-01	-3.684136	-0.003072	0.000000	0.000000e+00	7.213174e+00
ead_growth	22331.0	-1.613326e-02	4.601632e-01	-7.213263	-0.009916	0.000000	0.000000e+00	6.982210e+00
expected_loss_growth	22331.0	-1.349663e-02	3.009425e-01	-4.992291	-0.020175	0.000000	5.876101e-04	4.307961e+00
external_triggers_counter	22331.0	5.754652e-03	4.616390e-02	0.000000	0.000000	0.000000	0.000000e+00	1.916667e+00
internal_triggers_counter	22331.0	5.203621e-01	6.961988e-01	0.000000	0.083333	0.333333	7.500000e-01	1.425000e+01
negative_articles	22331.0	1.398202e-02	2.237657e-01	0.000000	0.000000	0.000000	0.000000e+00	1.658333e+01
outstanding_growth	22331.0	-1.184246e-02	4.622053e-01	-4.424374	-0.010822	0.000000	0.000000e+00	5.858981e+00
positive_articles	22331.0	1.102904e-02	1.553571e-01	0.000000	0.000000	0.000000	0.000000e+00	1.333333e+01
rwa_growth	22331.0	-2.071334e-02	4.420253e-01	-7.091768	-0.021432	0.000000	7.847310e-04	6.520162e+00
tot_allocated_limit	22331.0	3.283909e+07	7.259852e+08	0.000000	5000.000000	500000.000000	1.500000e+07	1.002000e+11
tot_outstanding	22331.0	1.646684e+07	4.965295e+08	0.000000	0.000000	4909.120000	4.558612e+06	7.071330e+10

Figure 5.2: Statistical data of growth and triggers-related features

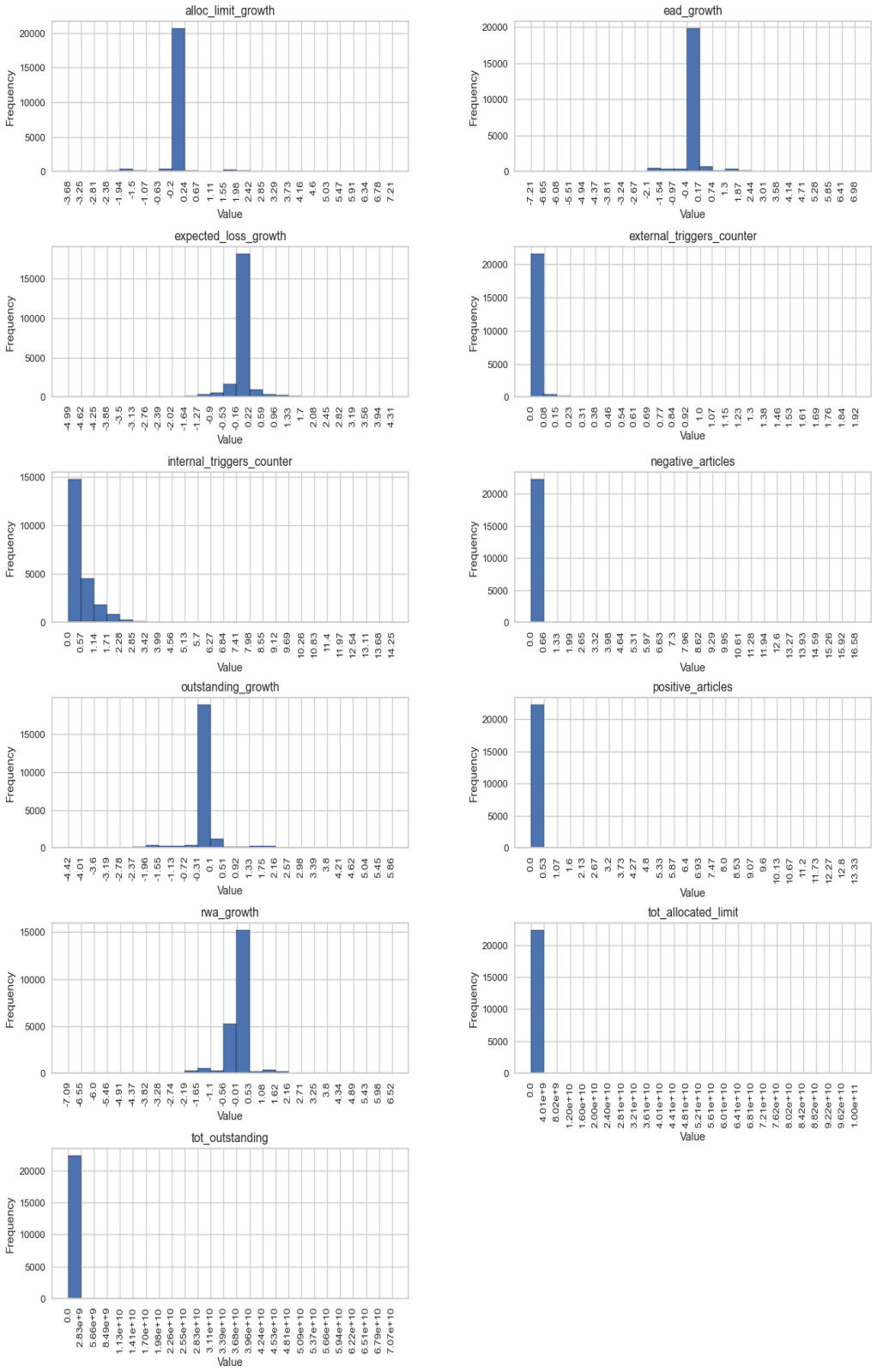


Figure 5.3: Clients distribution for different values of the growth and triggers-related features

5.2. DIMENSIONALITY REDUCTION RESULTS

5.2.1. DIMENSIONALITY REDUCTION: CORRELATION ANALYSIS

In the context of dimensionality reduction, correlation analysis can play a pivotal role in determining and quantifying the relationship between the different variables (Section 4.4.1). This technique can be leveraged in order to select the features that retain the most information and reduce the computational complexity of the model for better interpretations and visualisations of the clusters obtained. Thus, Figure 5.4 illustrates the correlation matrix that was developed on the full dataset, reporting the correlation score computed for every tuple of attributes included in the study. To enhance visibility and facilitate the identification of strong correlations, a gradient color scheme was applied across the matrix, as indicated by the color bar on the right. For features presenting a positive linear correlation, the respective cell would exhibit a red-toned colour whose intensity corresponded to the strength of the correlation between the two variables. On the contrary, features presenting a negative linear correlation were visualised using different shades of the colour blue instead. Furthermore, within this study, all the attributes presenting a correlation score exceeding 0.95 or below -0.95 were classified as highly correlated.

Figure 5.4 demonstrated that all the significantly correlated attributes displayed substantial positive correlations, some even as high as 0.99 for specific pairs. Specifically, the following pairs presented a correlation score above 0.95:

- RWA growth and Outstanding growth;
- EAD growth and Outstanding growth;
- RWA growth and EAD growth;
- Total Outstanding Amount and Total Allocated Limit;

In the case of the first three attribute tuples, the strong correlation was motivated by the underlying calculation implemented to compute these metrics. In fact, it was discovered that the outstanding amount represented a driving variable for the calculation of both the EAD and the RWA, thereby explaining the robust correlations observed. Nevertheless, despite the potential impact that numerous attributes can have on the clustering model's performance and based on the feedback and knowledge shared by credit risk experts of the bank, the researchers' discerned that the comparison of the variations that these three variables manifested throughout the study would yield valuable insights into clients' evolving risk exposure over the months. For this reason, even though they indeed exceed the pre-defined correlation threshold, none of the three attributes in question was discarded from the original dataset.

For what concerned the last tuple of features, namely the Total Outstanding Amount feature and Total Allocated Limit feature, a different conclusion was derived. Since the primary aim of the current study focused on the identification of an ideal cluster of entities, potentially eligible for future credit limit extensions, it was believed that preserving the information related to the most recent value of the allocated limit would have allowed the clustering algorithm

to generate segments of entities sharing similar amounts of credit limits. This would enable the stakeholders to re-define their strategies and offer more targeted solutions and incentives based on the current limit that characterises each cluster. Therefore, it was decided to retain the Total Allocated Limit attribute, while excluding the Total Outstanding Amount feature from the study. In summary, by removing the Total Outstanding Amount variable, the final dataset was streamlined to comprise only 19 out of the 20 initial attributes.

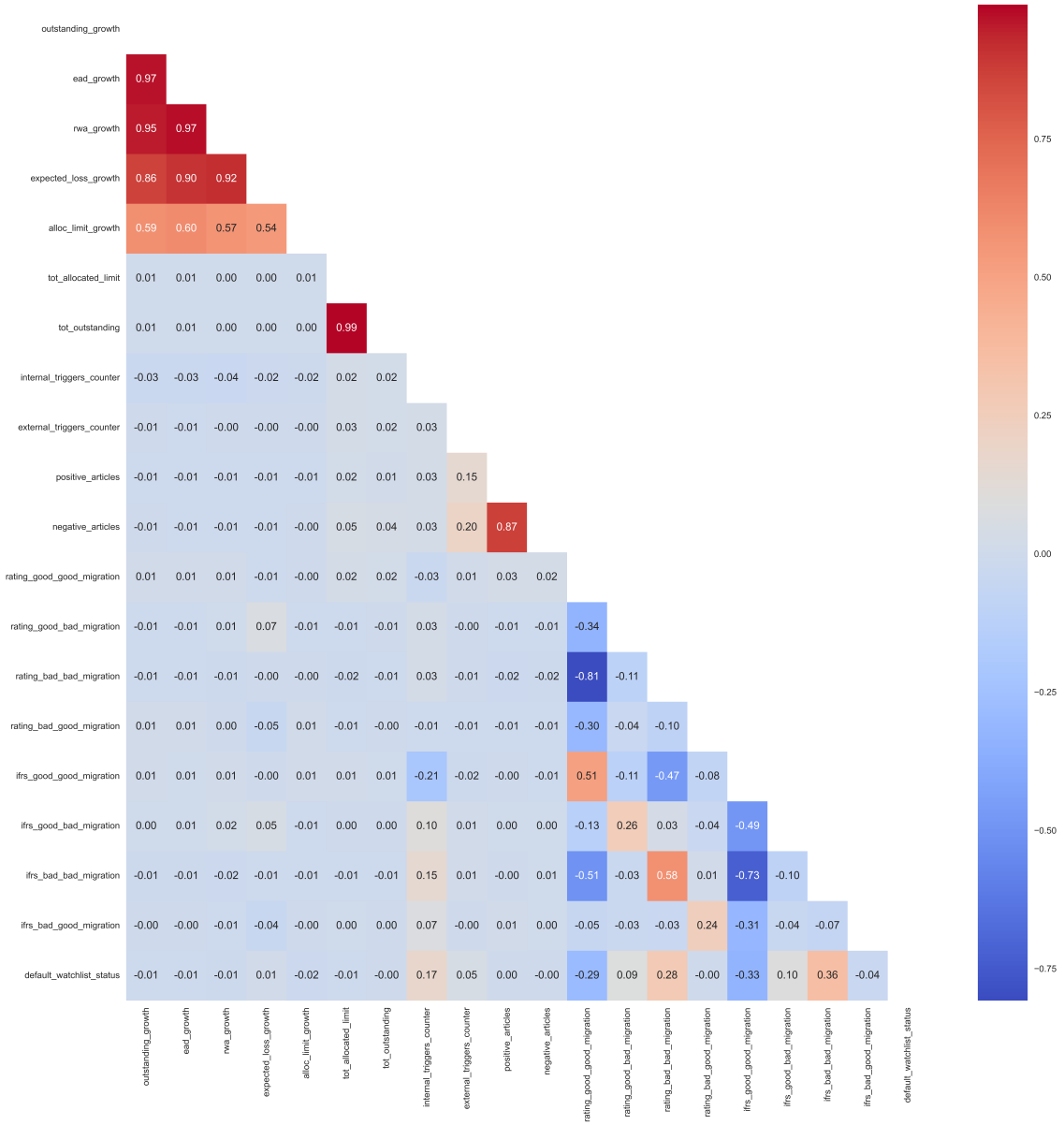


Figure 5.4: Feature's correlation matrix

5.2.2. DIMENSIONALITY REDUCTION: PCA

The purpose of utilising PCA as a secondary dimensionality reduction technique was to compare the efficiency and performance of the clustering algorithms for different dimensionality-transformed datasets, in order to determine the most appropriate technique and approach to

use for this specific set of data. Therefore, the cumulative variance plot, depicted in Figure 5.5, was generated and examined to support the decision of how many PCs should have been retained. As already mentioned in Chapter 4.4.1, the explained variance criteria that was deployed to select the appropriate number of components aided at retaining an explained variance ratio ranging between 80% and 95%. From the analysis of the aforementioned plot, it was discovered that the first 11 components, indeed, retained, approximately, 90% of the initial information and were, therefore, regarded as the optimal number of PC's. Then, having selected the ideal PC's for this specific study, the dataset was transformed into a reduced-dimensional space, where each feature's value of every data point was represented along the 11 PC's.

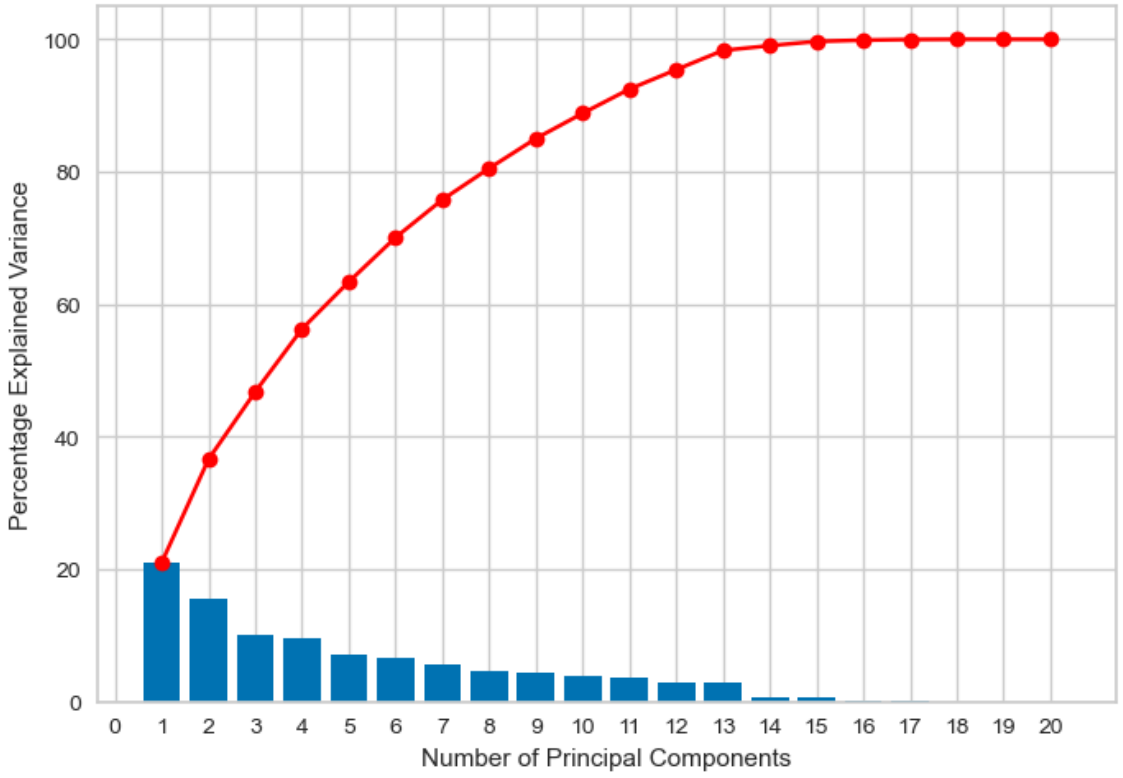


Figure 5.5: Cumulative variance for each component

5.3. MODELS IMPLEMENTATION RESULTS

5.3.1. K-MEANS: DETERMINING THE NUMBER OF CLUSTERS

In this section, the outcomes of the implementation of the Elbow Method technique, used to select the appropriate number of clusters for the K-Means algorithm, are presented for both the datasets obtained from the application of the two different dimensionality reduction techniques. The identification of the respective inflection point, known as the "elbow point", on the WCSS curve was supported by the deployment of the KElbowVisualizer tool [1]. Indeed, the tool was capable of automating the process of fitting the K-Means model for a number of clusters, previously defined by the user, and detecting the corresponding elbow point.

[1] https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

Figure 5.6 and Figure 5.7 provide an overview of the results obtained from the deployment of Python's Yellowbrick library [2] for the different datasets obtained. The use of the library in question proved to be particularly convenient and helpful since, in both portrayed scenarios, the inflection point was not as evident and noticeable as expected. In the case of the dataset generated from correlation analysis method (Figure 5.6), it was observed that, by selecting 16 clusters as the maximum number of segments, the inflection point identified by the tool, annotated by the vertical dotted line, aligned with, respectively, 9 clusters and reported an average sum of squared distances equivalent to 181101.883. For what concerned the PCA-transformed dataset, shown in Figure 5.7, the optimal number of clusters discovered by the tool was equivalent to 8 segments and the corresponding average sum of squared distances value was reportedly 171236.431.

In comparison to what had emerged during the review of the literature related to similar studies, the results collected in this particular research were relatively unusual and slightly out of the norm. In fact, if in most of the use cases examined the appropriate number of customers clusters would range between 4 and 7 segments [17, 24, 26], in the current investigation the groups generated were visibly more significant. A potential motivation for this specific behaviour could be associated to the utilisation of an equally noteworthy number of attributes and features, which led to the creation of detailed and elaborated customer profiles and more precise and accurate segments.

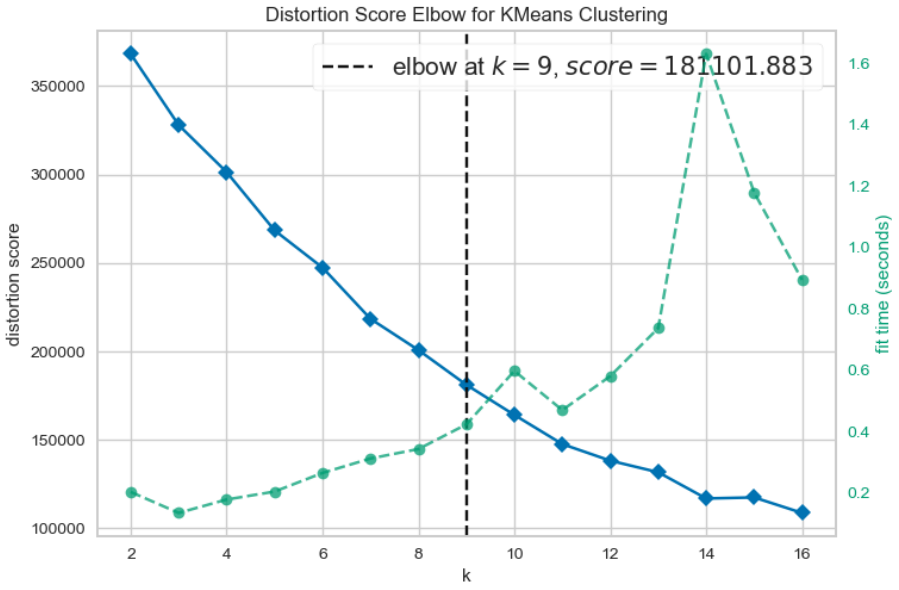


Figure 5.6: Elbow point detected for the dataset generated from correlation analysis

[2] https://www.scikit-yb.org

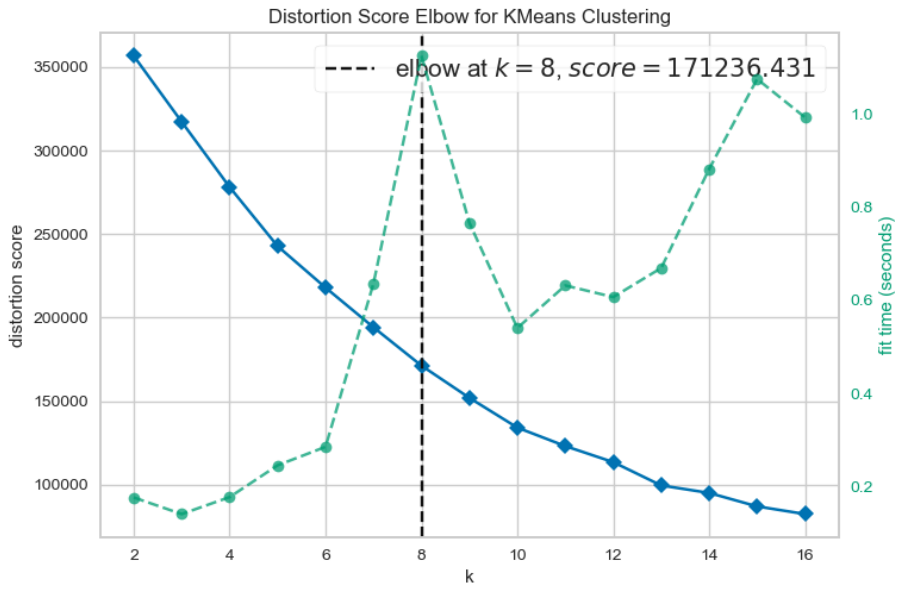


Figure 5.7: Elbow point detected for the dataset generated from PCA analysis

5.3.2. DBSCAN: DETERMINING ϵ AND THE MINIMAL NUMBER OF POINTS

As already outlined in the previous chapter, none of the conventional rules of thumb typically employed by researchers to determine optimal value of DBSCAN hyperparameters proved to be adequate for the type of datatset under examination. Consequently, a novel approach was devised, focusing on the optimisation of the number of clusters and the corresponding clusters Silhouette score.

Figure 5.8, Figure 5.9 and Figure 5.10, Figure 5.11 illustrate that the two pair of matrices, showcasing the respective number of clusters and the Silhouette scores obtained for different values of ϵ and MinPoints concerning the two datasets created (described in Section 3.3.3). Through the analysis of these matrices, it was observed that, in both cases, for increasing values of both hyperparameters, the corresponding computed Silhouette coefficient reported to have more favourable and significant scores. On the contrary, by incrementing the values of ϵ and the MinPoints hyperparameter, a decreasing number of clusters was produced by the DBSCAN model. Nevertheless, it was also noticed that for a number of segments somewhat equivalent to those generated using the K-Means algorithm, the corresponding value of the quality coefficient appeared to be rather reasonable and suitable for the study. Therefore, for both the uncorrelated and PCA datasets, it was established that the creation of 8 clusters with a respective Silhouette Score of, approximately, 0.47 represented the optimal trade-off between the two variables and defined an appropriate balance between the granularity of clusters and their quality.

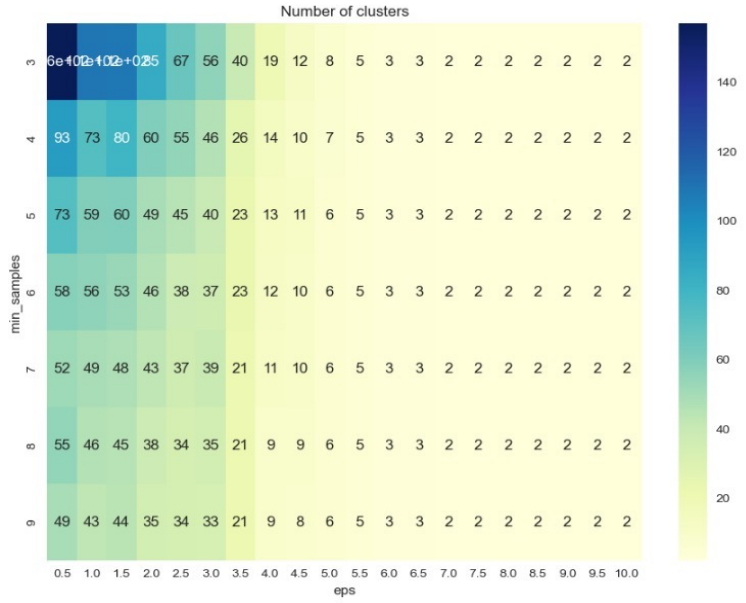


Figure 5.8: Number of clusters obtained for different DBSCAN hyperparameter values using the dataset generated from correlation analysis

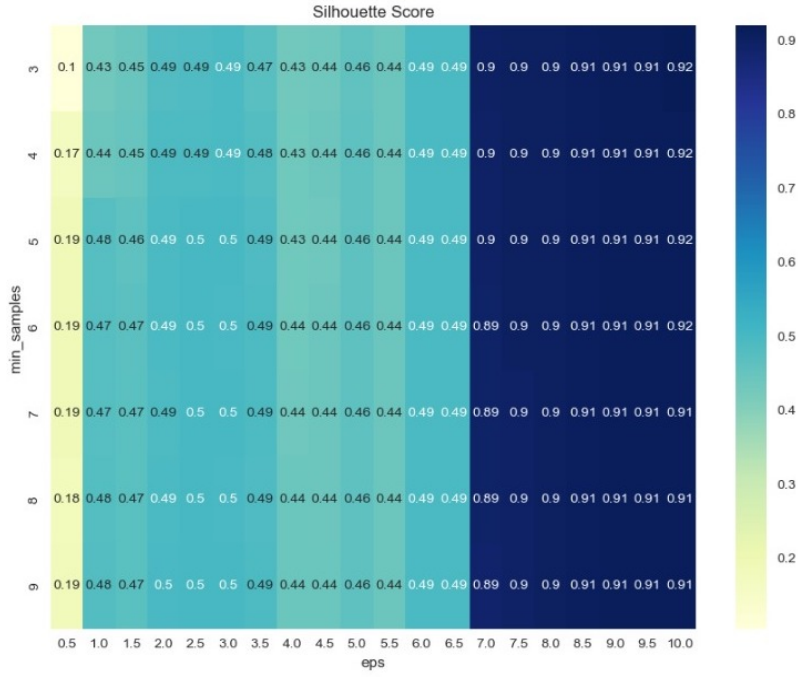


Figure 5.9: Silhouette scores obtained for different DBSCAN hyperparameter values using the dataset generated from correlation analysis

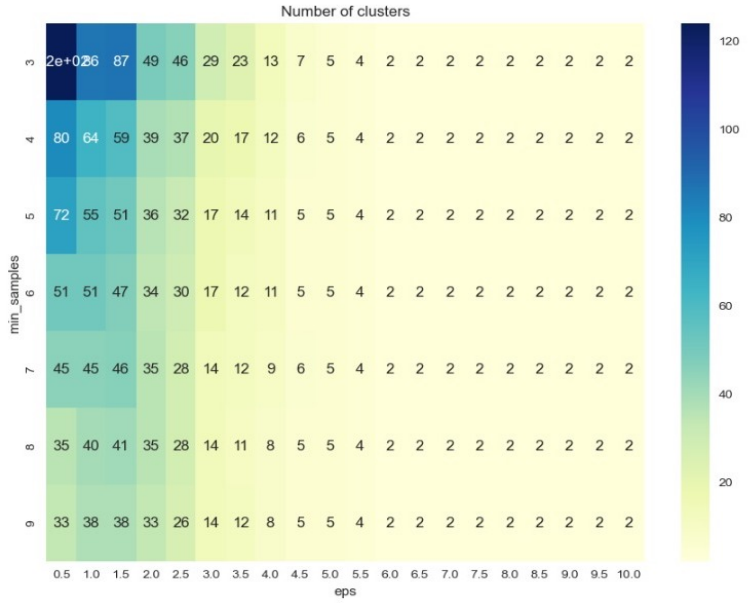


Figure 5.10: Number of clusters obtained for different DBSCAN hyperparameter values using the dataset generated from PCA analysis

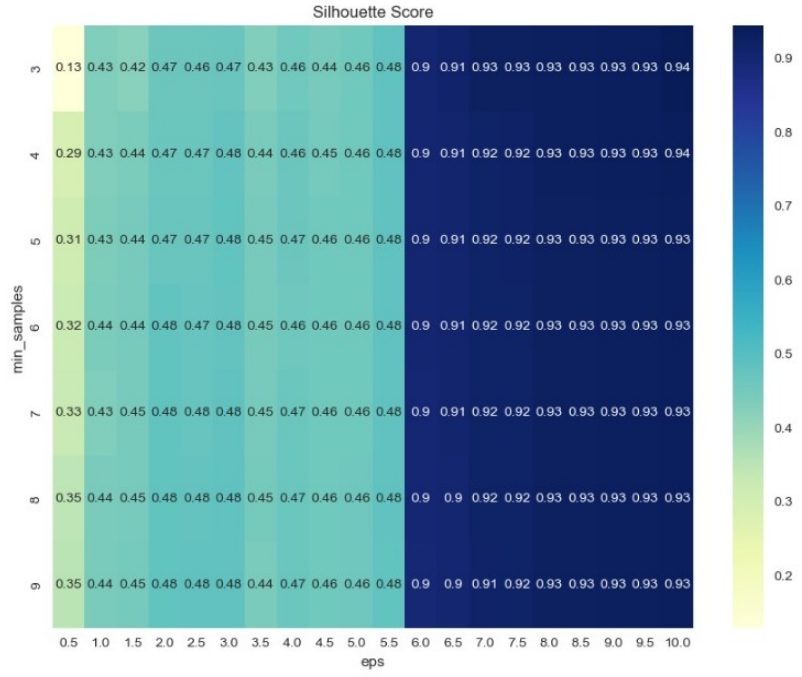


Figure 5.11: Silhouette scores obtained for different DBSCAN hyperparameter values using the dataset generated from PCA analysis

5.4. MODELS PERFORMANCE EVALUATION

Clustering can be defined as inherently an exploratory technique that, unlike supervised models, lacks of ground-truth labels that would allow the user to properly evaluate its performance. However, through the calculation of several metrics, capturing different aspects of the cluster-

ing quality, a more comprehensive understanding of how well the model performs and its effectiveness can be achieved. In this study, the indicators that were computed to assess the clustering robustness were namely the Silhouette Score, the Davies-Bouldin Score and the Calinski-Harabasz Score (Section 4.5.1). These metrics were implemented in order to determine the intra-cluster compactness and inter-cluster separation for both K-Means and DBSCAN. An overview of the results obtained is provided in Table 5.1. One significant finding from the analysis of these metrics related to the variation in results obtained using the same clustering algorithm on different datasets. Concerning the Silhouette Score, since the values for the uncorrelated and the PCA dataset were respectively equivalent to 0.538135 and 0.538501 in the case of K-Means and 0.458367 and 0.468183 for DBSCAN, a minimal contrast was observed for both models. Nevertheless, it was also revealed that, under K-Means, the PCA-transformed datasets demonstrated relatively enhanced performance for both the Davies-Bouldin and Calinski-Harabasz metrics, reportedly equal to 0.759579 and 4503.09, and, thus, generating more compact and distinct clusters. This trend was similarly noticed also when implementing the DBSCAN clustering algorithm, with the exception of the Davies-Bouldin score, which proved to be slightly more significant for the uncorrelated dataset instead. Still, the utilisation of a dataset subjected to a conversion to its PCs can impose several limitations from a practical perspective. For instance, since PCs represent linear combinations of the initial attributes, they may not have a clear and intuitive meaning to managers, who are inherently disconnected from the original features and unable to directly interpret each feature's contribution to the final outcomes. On the other hand, traditional feature selection techniques, such as correlation analysis, do require more computational resources and domain knowledge regarding variables' inter-dependencies in order to detect actual information redundancy.
In addition, the table showed that, for each one of the performance indicators exploited, the values reported for K-Means were remarkably more promising and meaningful, suggesting that the this particular algorithm produced tighter and better defined clusters. For instance, in the case of K-Means applied to the dataset formed over the correlation analysis process, the reported Calinski-Harabasz score, equivalent to 3746.80, significantly exceeded the score obtained for the same dataset using the DBSCAN model, equal to 859.201.

Table 5.1: Clusters quality results for different algorithms and different datatsets

Dataset	Algorithm	**Silhouette**	**Davies-Bouldin**	**Calinski-Harabasz**
Uncorrelated	K-Means	0.538135	0.814055	3746.80
	DBSCAN	0.458367	1.56887	859.201
PCA	K-Means	0.538501	0.759579	4503.09
	DBSCAN	0.468183	1.86492	1183.54

5.5. SEGMENTS EXPLORATION

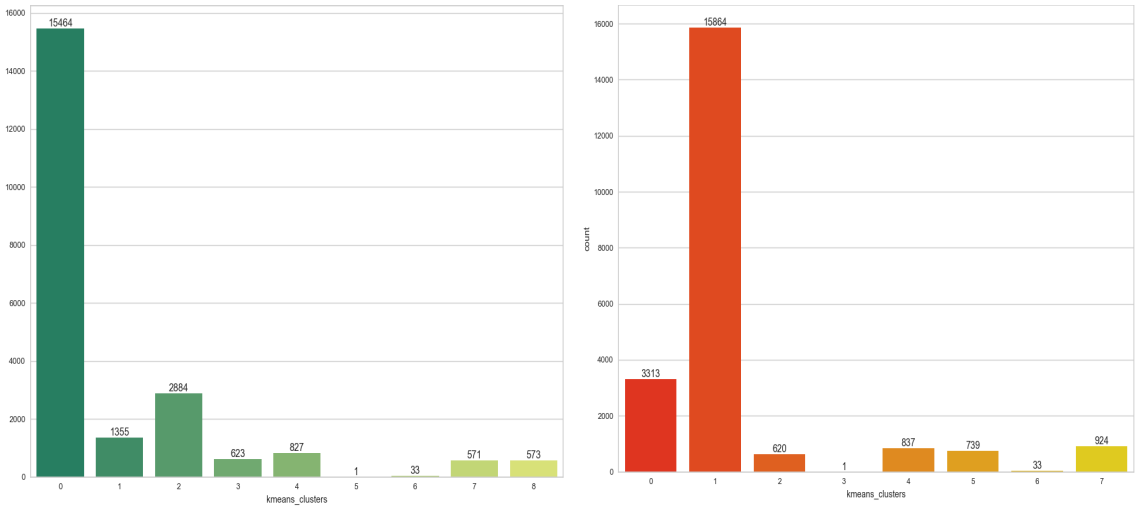
The main purpose of this subsection is to visualise and investigate the results related to the implementation of the clustering algorithms over the different datasets obtained from the di-

mensionlity reduction procedures. As outlined in Chapter 4.5.2, this subsection is divided into a number of paragraphs that present and discuss different aspects of the clusters exploration. However, because of the poor results that the DBSCAN model presented during the models' performance evaluation process, the in-depth analysis of the actual segments generated from the model's deployment is omitted from this chapter. Nevertheless, an overview of the descriptive statistics examination is provided in the Appendix C and illustrated in Figure C.14, Figure C.15 and Figure C.16. Indeed, from the study of these charts, it was revealed that the algorithm proved to be rather inefficient at identifying underlying patterns and common traits among the client base, as none of the clusters generated presented any particular characteristic or property that distinguished them from the other segments, unlike the ones obtained from the utilisation of K-Means. Therefore, it was confirmed that the no valid and conclusive insight could be derived from the exploration of these subsets of clients.

5.5.1. CLUSTERS' DENSITIES ANALYSIS

One fundamental aspect in order to assess the performance and effectiveness of the clustering algorithm relies on the analysis of the clusters' densities and sizes. This specific investigation provided significant insights regarding the actual compactness and separation of the segments and revealed meaningful information about the outliers and noise points found in the dataset. Figure 5.12 represents an overview of the clients' distributions across different clusters generated from the implementation of the K-Means clustering algorithm on top of the datasets obtained from the application of the dimensionality reduction procedures. From the investigation of these very charts, it was discovered that the clusters formed using the two individual datasets defined the same subset of entities and portrayed an identical scenario under different cluster labels. Indeed, it appeared that, in both cases, the segments developed were visibly heterogeneous and characterised by critical variations in terms of density. Nevertheless, the presence of sparse and low-density clusters did not necessarily entail that the algorithm struggled at distinguishing the data points from each other. In fact, depending on the nature of the data, it may occur that valid interpretations can be derived from a more in-depth exploration of the segments generated.

In addition, it was also observed that one particular cluster was composed by approximately 70% of the overall client base considered for this study and two minor subgroups were characterised by a limited and negligible sizes of, respectively, 1 and 33 entities. Finally, the remaining segments presented more even and less sparse distributions, suggesting a more effective and distinct separation of the clients involved.



(a) Distribution for the uncorrelated dataset

(b) Distribution for the PCA dataset

Figure 5.12: Clients distribution across different K-Means clusters

5.5.2. CLUSTERS DESCRIPTIVE STATISTICS

UNCORRELATED DATASET: K-MEANS

Through a higher-level examination of the distribution and averages for each feature within the generated clusters, depicted in Figure C.1 and Figure C.2, distinct patterns and traits were readily identified. Firstly, with regards to the "default_watchlist_status" attribute, it was observed that, for three specific clusters, namely Cluster 1, Cluster 2 and Cluster 6, over 10% of the entities included were assigned to a critical status in April 2023, indicating a high a default risk for the borrowers involved. The remaining segments, instead, reported a less concerning scenario, with fewer than 5% of distressed or watchlisted customers. For what concerned the variables linked to monthly growths, it was determined that, for most segments, these particular features did not contribute in a significant manner to the clustering and grouping of the entities. This conclusion was drawn from the detection of a substantial number of outliers and noise data points that these attributes presented in each cluster. Indeed, only Cluster 3 and Cluster 4 appeared to be composed by borrowers that shared similar and notable growths, as demonstrated by the boxplots depicted in Figure C.1. In the case of the former segment, the positive growths reported defined consistent and aligned increases across all the the five features in question, potentially due to the initiation of a new loan. Cluster 4, instead, showcased remarkable decreases and negative growths, likely caused by a full repayment of the clients' open loan. Nevertheless, for both subsets of customers, it was noticed that, in terms of the allocated limit growth, an elevated number of clients involved displayed either null or opposite growths compared to the entities falling into the feature's interquartile range. This suggested a minor significance and contribution for the "allocated_limit_growth" attribute. As for the other clusters, despite the number of outliers reporting unusual values, most of the entities involved were characterised by more ordinary and modest monthly increments and decrements.

Similarly to what emerged for the growth-related features, the analysis of the attributes related

to the average monthly triggers and news articles also demonstrated limited importance in the clusters generation process. In fact, not only most segments included a noteworthy percentage of outliers, but also the values reported for the largest portion of clients were visibly coincident and overlapping across the different clusters. This phenomenon was particularly pronounced for the "external_triggers_counter", "negative_articles" and "positive_articles" features, which were, reportedly, equivalent to 0 for the majority of the borrowers assigned to each segment, with the exception of Cluster 6, indicating a group that activated a concerning number of triggers and news articles on a monthly basis. In the case of the "internal_triggers_counter" attribute, although there was no real distinction among the different clusters, it was still observed that Cluster 2, Cluster 6 and Cluster 7 comprised more borrowers characterised by a higher likelihood of generating one or more triggers every month.
Regarding the credit status migration features, the clusters separation was deemed to be more conspicuous and defined for this subset of attributes, as a greater portion of entities included in each segment shared concordant migrations and behaviours related to their credit risk situation. For instance, it was discovered that entities that preserved a favourable and promising credit quality throughout the whole activity period would be most likely assigned to either Cluster 0, Cluster 3 or Cluster 4. On the contrary, Cluster 2 and Cluster 1 were composed by clients that either exhibited a negative migration and a depreciation of their credit quality, or that were characterised by more critical and high-risk credit scores both at the beginning and at the end of the study. Finally, from the exploration of the descriptive statics measures, it was discerned that the "tot_allocated_limit" attribute played a pivotal role only for the identification and separation of the one individual entity involved in Cluster 5, which reported an outstanding and remarkable limit compared to the other clusters.

Nevertheless, in order to obtain a more detailed representation and a better understanding of the scenario that each cluster displayed, an in-depth statistical analysis of the main patterns and characteristics defining each subgroup of client was implemented. The results of such investigation are outlined in the following subsection:

- **Cluster 0**:

    1. *Default/Watchlist Status*: As of April 2023, the stem plot chart (Figure C.2 seemed to indicate that most of the clients included were not being watchlisted or facing a critical default scenario, except for a minimal percentage of entities.

    2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: Concerning the aspects of monthly growth, the descriptive statistics values attributed to the majority of clients involved in this set of entities were notably minor and lacking in significance. However, the 25^{th} percentile values for each feature, shown in Figure C.4, suggested that a limited group of clients showed minor declines in the respective feature. Nevertheless, these decrement did not surpass the -1% for any of the variables in question. In addition, although this clusters contained a considerable number of instances deviating from the norm across these metrics, the

limited decreases and increases were also observable in the value displayed in the stem plots.

3. *Average number of monthly triggers and news articles*: Regarding the external triggers, the majority of the clients presented a number of monthly external triggers equal to 0, as indicated in Figure C.4. Nevertheless, the feature's box plot chart (Figure C.1) demonstrated that, for the clients that were not assigned with values included within the interquartile interval, also labeled as outliers, the recorded average would be greater than 0, with a maximum of almost 2 external triggers per month. In terms of monthly internal triggers, instead,the range fell between, approximately, 0 and 0.6 for most entities. Furthermore, as a whole, this group typically did not have any positive or negative monthly article detected on a monthly basis, excluding the outlier clients.

4. *Worst IFRS stage and Internal Rating score migrations*: It seemed that, with respect to both the Internal Rating score and Worst IFRS stage shifts, the majority of the included entities retained a "good" levels from the outset. However, it is worth mentioning that, in terms of Internal Rating, approximately $10-15\%$ of the borrowers either experienced a downward migration or were assigned with poor scores from the beginning (Figure C.2).

- **Cluster 1**:

1. *Default/Watchlist Status*: Most of the clients included were associated with a regular status in April 2023. However, for this cluster a 16% of the customers, reported in Figure C.4), were labeled either as in default or being watchlisted and thoroughly financially monitored.

2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: In a manner similar to the situation examined in the preceding cluster, these customers commonly exhibited substantial or minor negative log-linear growths for these features. On the other hand, the 25^{th} and 75^{th} percentiles of Figure C.4 seemed to indicate that this cluster encompassed, respectively, more clients that showed more significant monthly decreases and pronounced monthly increases in the Expected Loss and RWA compared to their corresponding Outstanding Amount and EAD. Concerning the anomalies observed in relation to these attributes, similarly to the other clusters, there were notable instances of outliers within this specific segment (Figure C.3).

3. *Average number of monthly triggers and news articles*: Much like Cluster 0, these entities, on average, experienced no external trigger being raised on a monthly basis. Even among the exceptional cases, the highest recorded average of the number of monthly external triggers remained below 0.8 (Figure C.1). As for the average of monthly internal triggers, instead, the situation appeared to be more noteworthy in

comparison to Cluster 0. In fact, the reported values were slightly more elevated, with the 75^{th} percentile exceeding the value of 1 and a maximum of 7.5 for the outliers (Figure C.4). Furthermore, the statistics indicated that for most clients, neither positive nor negative articles would be flagged within each month, excluding the customers labeled as outliers.

4. *Worst IFRS stage and Internal Rating score migrations*: While the majority of the entities were categorised under a relatively low-risk activity status in April 2023, the analysis of the credit quality migrations revealed a consistent trend: in terms of Worst IFRS stage, all clients exhibited a negative migration (Figure C.2). In contrast, this migration did not manifest correspondingly in the context of the Internal Rating standard. Specifically, although a substantial number of clients preserved their initial “good" rating score, certain entities encountered either a negative migration or were designated as low-credit-quality borrowers from their starting month.

- **Cluster 2**:

1. *Default/Watchlist Status*: Even though most clients included were not assigned to high-risk status in the last month of the study, over 30% of the entities were actually being watchlisted or dealing with in a default situation (Figure C.4).

2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: The findings of the examination of the descriptive statistics related to the the monthly growths of the EAD, RWA, Expected Loss, Allocated Limit, and Outstanding Amount features, reported in Figure C.4, demonstrated the presence of an alignment among the values. Notably, a predominant portion of the observed values was confined within a narrow range around zero. These positive and negative fluctuations indicated that this cluster contained entities that presented both of increments and decrements for the respective features. Moreover, it is crucial to acknowledge the prominence of outliers, which were present in notable numbers across all four attributes.
   Finally, it is also of significant importance to stress the fact that, similarly to the previous clusters, this group of entities was not characterised by any augmentation in their overall allocated credit limit. Instead, since a minor subset of entities actually presented a decrease in this attribute's growth, it is believed that they either encountered a reduction in their available limit or engaged in new loan agreements entailing lower limit.

3. *Average number of monthly triggers and news articles*: Concerning the average count of external triggers, the data presented the same prevailing pattern (Figure C.4): nearly all clients were associated with, approximately, 0 external monthly triggers, and even among the exceptional cases, the maximum average remained below 0.8. For what concerns the internal triggers, it was also discovered that, based on the

values of the 25^{th} percentile and the median, more entities with a minor likelihood of generating 1 internal trigger per month were included compared to the previous clusters. However, the 75^{th} percentile also suggested that more clients raising over 1 internal trigger each month were involved as well. The examination of the average number of monthly positive and negative news articles showed that, once again, most customers were not detected in any article every month. While the outliers within the cluster did display noteworthy quantities of both positive and negative articles (Figure C.1, these elements exerted a minimal influence on the overall findings, representing a mere 5% of the total entity count.

4. *Worst IFRS stage and Internal Rating score migrations*: Despite a relatively minimal number of exceptional clients, almost all the entities involved were characterised by no score or stage migration. Instead, they were all associated with scores above the threshold levels for both features (Figure C.2).

- **Cluster 3**:

1. *Default/Watchlist Status*: Regardless of the 4% of defaulted or watchlisted entities dealing with loan repayment difficulties, shown in Figure C.4, the cluster was mainly composed of entities in regular statuses.

2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: For what concerns the growth-related features, this segment comprised borrowers who encountered significant monthly increments across all five attributes. Indeed, it was noticed that for each monthly growth feature, with the exception of the Allocated Limit and the Expected Loss attributes, the reported increase generally fell within the range of 1.3 and 1.9 for most entities (Figure C.4). This trend was likely due to the fact that, in their initial month, most of the entities in question presented null values for each feature. Moreover, it is important to emphasise that, although the considerable magnitude of these growths, some of the clients involved may have actually witnessed a reduction of their risk exposure. This assumption was supported by the disparity between the growth reported in all the statistical metrics for the Expected Loss and the EAD feature.

3. *Average number of monthly triggers and news articles*: The average number of monthly internal triggers seemed to remain rather modest and minimal compared to the previous clusters, with the 75^{th} percentile falling below the value of 0.7 (Figure C.4). This pattern persisted even among the feature's outliers cases, for which the reported statistical values also proved to be insignificant. Similarly, the statistics associated with the number of monthly external triggers did not suggest any substantial risk or a concerning scenario.

4. *Worst IFRS stage and Internal Rating score migrations*: For both the Internal Rating Score migration and Worst IFRS stage standards, it was observed that most of

the entities included did not deal with any migration and were always associated with scores below the established thresholds. However, the cluster contained several instances for which a negative migration had occurred or that were labeled as “low-credit-quality" clients since their first month of activity (Figure C.2).

- **Cluster 4**:

  1. *Default/Watchlist Status*: Compared to Cluster 3, this segment appeared to contain a higher number of clients assigned with “Default" or “Watchlist" status in April 2023. This subgroup of entities, however, was still considered rather limited and not concerning.

  2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: When examining the growth-related features, the situation portrayed for this specific category seemed to be completely contrary to the situation analysed for the previous cluster. In fact, as indicated by the boxplots, all the borrowers involved experienced with significant decreases in all five measures. These decrements also appeared to align with one another in terms of magnitude (Figure C.4). However, for the Expected Loss and Total Allocated Limit growths the declines were not as steep and intense, since they also involved entities for which no negative growth was identified. Nevertheless, it was believed that these substantial drops were caused by a resetting of the features' values in the last month recorded for the entity, which led to a negative logarithmic growth below the value of -1.

  3. *Average number of monthly triggers and news articles*: Regarding the external triggers, the charts indicated that this set of customers presented a notable low average of monthly external triggers, even when considering the outliers within the cluster. A comparable observation was derived from the news articles. For what concerns the average monthly internal triggers number, it could be affirmed that this group of borrowers would not raise a substantial number of internal triggers every month. In fact, the box plot of Figure C.1 illustrated that the 75^{th} percentile fell below the value of 1 for this attribute, indicating that 75% of the entities in question would not generate an internal trigger every single month. Still, the box plot also revealed that this group includes instances where the average monthly triggers count could reach as high as 11 triggers.

  4. *Worst IFRS stage and Internal Rating score migrations*: As for the Worst IFRS stage and the Internal Rating score migrations, it was observed that, for both attributes, most of the clients were associated with “good” credit qualities both at the start and at the end of their activities. Nonetheless, there were several cases where the clients experienced either a downward migration or were consistently categorised with negative scores.

- **Cluster 5**:

1. *Default/Watchlist Status*: This cluster contained one single entity which, as of April 2023, was not facing a distress situation or involved in any watchlisting process (Figure C.1).

2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: For this particular entity, there was a minor log-linear increase of, approximately, 0.025 for both the EAD and the Outstanding Amount features, which were probably derived from an even more significant increase of the Allocated Limit attribute (Figure C.1). These increases did not affect the client's RWA or the Expected Loss, that, instead, saw a marked negative growth over time. This phenomenon could be have been caused by a potential decrease of the risk exposure- associated with this specific borrower or a depreciation of its PD or LGD.

3. *Average number of monthly triggers and news articles*: Concerning the average number of monthly external triggers, it appeared that this entity generated, approximately, zero external triggers on a monthly basis. However, a different scenario emerged from the analysis of the internal triggers average. Indeed, the statistics (Figure C.4) indicated that this entity presented an average of 0.8 internal triggers per month, suggesting that one single internal trigger would be flagged almost every month. Moreover, the study of the number of positive and negative news articles indicated that this borrowers would be detected in, at least, one negative every month.

4. *Worst IFRS stage and Internal Rating score migrations*: With regards to the Worst IFRS Stage stage and Internal Rating score, the entity in question was not subjected nor to a positive or a negative migration over time. On the contrary, the borrower was always associated with positive credit-quality scores throughout the whole study.

5. *Most recent allocated limit*: The credit limit that was allocated to this major entity in April 2023 was significantly higher compared to the entities included in the other clusters, as highlighted in Figure C.1, revealing that the client in question may have represented a financially-healthy and wealthy corporate.

- **Cluster 6**:

1. *Default/Watchlist Status*: The percentage of clients that, in April 2023, were associated with a critical status was greater than 12% (Figure C.4, which was remarkably high compared to the other segments.

2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: Upon examining the growth for the Outstanding Amount attribute, reported in Figure C.4, it was noticed that, for the majority of the clients, the values spanned from -0.1 to 0.064, centering around a median of 0. This indicated that approximately 50% of the entities involved exhibited negative growth rates for this specific feature. However, the remaining 50% of these customers experienced notably positive

growth rates instead. This pattern was not as pronounced in relation to the other attributes, where nearly almost all statistical metrics, including medians and percentiles, reported mainly negative decreases, for the Expect Loss, the RWA and the Outstanding Amount in particular, or no general growth.

3. *Average number of monthly triggers and news articles*: Based on the investigation of the average number of monthly external triggers, as underscored by the boxplot C.1, it appeared that, while half of the entities would not generate monthly any external trigger, the remaining 50% were more likely to raise one single external trigger, as reported by the attribute's 75^{th} percentile. Additionally, it was noticed that, in comparison to the preceding cluster, these entities also exhibited a greater frequency of monthly internal triggers. In fact, the feature's median value was, approximately, equivalent to 1, hinting that nearly all the included clients produced one internal trigger each month.
   The observation was also visible in terms of monthly positive and negative articles as well.

4. *Worst IFRS stage and Internal Rating score migrations*: Despite the elevated averages of monthly triggers, most of the borrowers involved were associated with Internal Rating scores and Worst IFRS stages below the established thresholds, both in the starting and ending months of the study. Nevertheless, for both attributes, a remarkable number of exceptions was identified, concerning borrowers that faced bad or positive migrations or that, overall, were characterised by a consistent low credit quality.

- **Cluster 7**:

  1. *Default/Watchlist Status*: Approximately 99% (Figure C.4) of the entities involved were assigned with more regular and not critical status in April 2023 .

  2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: Concerning the EAD and Outstanding Amount growth features, the analysis of the statistical metrics of Figure C.4 indicated that these attributes shared a relatively consistent pattern and that minor declines occurred for nearly all clients within the cluster, although the Outstanding Amount displayed also a positive value for the 75^{th} percentile. Nevertheless, when studying the growths related to the RWA and Expect Loss features, a significant difference emerged compared to the aforementioned attributes. This distinction was particularly notable for the Expected Loss, as it appeared that a relevant number of clients experienced a more pronounced negative monthly growth on average compared to their corresponding EAD. This trend could be attributed to a reduction in the assigned values for PD or LGD for these entities. In terms of Allocated Limit, the data demonstrated that the majority of entities in question either dealt with a slight decrease in their allocated limit or

encountered no changes whatsoever.

3. *Average number of monthly triggers and news articles*: The values depicted in the plots pertaining to the average number of monthly external triggers (Figure C.2) did not appear to be overly alarming. In fact, it was observed that the overall mean for the external triggers was, once again, rather minimal and almost negligible. As for internal triggers, the average monthly number for these entities exceeded 0.8, indicating that the majority of entities within this cluster gathered around 1 internal trigger almost every month. This result, however, was partly influenced by the presence of 2.45% of outliers detected for this attribute (Figure C.3). Nevertheless, based on the results reported in Figure C.4 for the median and the percentiles, it seemed that the clients involved would, for the majority, raise either 1 or no internal trigger every month. Moreover, the statistics also demonstrated that for these class of entities, no positive or negative articles would be usually detected on a monthly basis.

4. *Worst IFRS stage and Internal Rating score migrations*: From the analysis of the Worst IFRS stage migrations, it was discerned that all borrowers underwent a transition from a “bad" classification to a Stage 1 of the standard, indicating a substantial enhancement in credit quality. This trend, however, was only partially reflected in terms of Internal Rating; it is worth noting that the majority of the entities were actually characterised by scores below the pre-defined threshold from the start of their starting month.

- **Cluster 8**:

1. *Default/Watchlist Status*: As of April 2023, approximately 9% of the clients assigned to Cluster 8 presented a critical financial situation or had been put in ING's WB clients watchlist (Figure C.4).

2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: The primary characteristics of this specific cluster revolved around the fact that, for all five attributes, it mostly comprised entities that did not manifest any growth over time. Specifically, both the 75^{th} percentiles and the medians related to these five features were, according to the statistics C.4, all equivalent to 0. The exceptions observed were slightly more significant in the context of Expected Loss and RWA, where the 75^{th} percentiles reported more marked and prominent monthly declines. In addition, it was discovered that, regardless of the 38.39% of outliers (Figure C.3), most clients did not obtain any extension or depreciation of their allocated limit over the months.

3. *Average number of monthly triggers and news articles*: For what concerns the average number of monthly external triggers, it was, once again, verified that majority of the customers involved would not be flagged by any external trigger each month, except for a limited number of extraordinary cases. Moreover, the distribution of

the values related to the average number of monthly internal triggers proved to be slightly less sparse and concerning compared to the previous clusters, as also depicted by the respective box plot of feature's distributions in Figure C.1. Indeed, the reported average number of monthly internal triggers was equivalent to, approximately 0.4, meaning that these entities were less likely to generate one internal trigger every month.

4. *Worst IFRS stage and Internal Rating score migrations*: Regarding the Internal Rating score migration, it was discerned that all the borrowers in question were subjected to a positive migration, symbolising an improvement of their credit quality. On the other hand, this migration was not detected in terms of Worst IFRS stage, where, instead, approximately 80% of the clients maintained their "good" levels and were assigned with Stage 1 throughout the whole study (Figure C.2).

PCA DATASET: K-MEANS

In the context of the clusters formed using the PCA-transformed dataset, the exploration of the clients' distribution within each cluster and for each feature demonstrated the recurrence of nearly identical patterns observed in the scenario of the uncorrelated dataset. Indeed, as already observed during the clusters density study of the previous subsection, it was reaffirmed that the entities involved in every segment had been clustered together on the basis of the same attributes that they shared in the previously analysed clusters. However, the segments in question were assigned with different labels and cluster rankings.

Nevertheless, in order to validate the authenticity of this high-level observation and presumption, a more articulated and comprehensive analysis of each feature's of statistical indicators was conducted across all segments. The subsequent section delves into a discussion of these findings in detail.

- **Cluster 0**:

  1. *Default/Watchlist Status*: Although the majority of the entities were not labeled as "In Default" or "Watchlist" in April 2023, it was observed that over 30% of the customers involved did manifest a critical and more serious condition (Figure C.7).

  2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: Despite the notable presence of a large number outliers for all five attributes (Figure C.8), this set of entities was mainly characterised by clients that did not show any increase for the Allocated Limit and EAD. In fact, the 25^{th} percentile and the 75^{th} percentile values reported in Figure C.6 were, reportedly, smaller and equivalent to 0. Despite this, it was also discovered that, for several elements, a notable rise of the Outstanding Amount, RWA and, in particular, the Expected Loss had actually occurred, as indicated by the attribute's 75^{th} percentile values. For this reason, it could be concluded that, because of the imbalance between the EAD growth and the Expected growth that some of the entities involved presented, a substantial number

borrowers may have dealt with an increase of their respective PD or LDG values.

3. *Average number of monthly triggers and news articles*: Based on the statistics and the plots of Figure C.7, the average number of monthly external triggers raised this group of borrowers was rather negligible. In contrast, the internal triggers seemed to have carried slightly more significance, with an average of 0.78 and a 75^{th} percentile exceeding the value 1. This implied that, a relevant number of entities with a high probability of one internal trigger each month was included. As for the negative and positive articles, it seemed that these that these clients did not typically attract substantial media attention. However, it is important to underline that, for each one of these attributes, this cluster contained a critical number of outlier instances that would activate a concerning amount of triggers every month.

4. *Worst IFRS stage and Internal Rating score migrations*: Regarding the Worst IFRS stage and Internal Rating score migrations, it was discerned that, based on both standards and the results illustrated in Figure C.7, the majority of the entities involved had been assigned to low-quality credit levels right from the outset.

- **Cluster 1**:

1. *Default/Watchlist Status*: In the case of Cluster 1, less than 1% of the borrowers involved were assigned to high-risk status in the most recent month of the study (Figure C.9).

2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: Based on the analysis of the growth-related features shown in Figure C.9, it was discovered that most of the entities included either showed a visible monthly decrease or no general growth was detected. Nevertheless, the statistics reported both by the box plots in Figure C.6 suggested that, for a number of clients, the decrement in the feature's value for the Expected Loss and RWA, were, in reality, particularly significant and evident compared to their corresponding EAD decrement. In fact, if the 25^{th} percentile values for the EAD and Outstanding Amount growth features were equivalent to -0.008, the reported values of the corresponding metric for the Expected Loss and RWA were, respectively, -0.01 and -0.02.

3. *Average number of monthly triggers and news articles*: Regarding the average number of monthly internal triggers, the situation that surfaced for this specific segment was less severe and threatening in comparison to earlier the cluster. This consideration was supported by the considerably lower average number of monthly internal triggers, spanning between 0 and 0.67 for most clients. A similar inference could be drawn when considering negative articles, as the number of monthly articles identified remained relatively modest, despite the several outliers.

4. *Worst IFRS stage and Internal Rating score migrations*: In relation to both attributes, the majority of borrowers (around 90% based on the results of Figure C.7) did not

experienced any changes in their score or stage; rather, they maintained a “good" credit level right from the first month of their activity, despite an irrelevant number of outliers and exceptions.

- **Cluster 2**:

  1. *Default/Watchlist Status*: The number of clients in default or that had been put in the company’s watchlist was slightly more significant than the previous cluster, comprising, approximately, the 3% of the total entities based on Figure C.9.

  2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: In contrast to the previous clusters’ examinations, these borrowers shared a common trait of experiencing significant positive growth across all five features. The expansions appeared notably pronounced for EAD, RWA, and Outstanding Amount, ranging from 1.3 to 1.9 in terms of logarithmic growth. Conversely, the statistical data indicated a less substantial rise in the Expected Loss and Allocated Limit for this group (Figure C.9).

  3. *Average number of monthly triggers and news articles*: As for the monthly external and the internal triggers collected, the data reported depicted in Figure C.7 an overall stability and lacks severity scenario. It was observed that, in fact, on average, no monthly external triggers were raised for this group of organisations and even for the customers labeled as outliers, one single external trigger would be activated every month. Moreover, while the 25^{th} percentile was somewhat higher compared to Cluster 1, the rest of the statistical measures reported in Figure C.9 for the monthly internal triggers aligned quite closely with those examined in the previous segment.

  4. *Worst IFRS stage and Internal Rating score migrations*: For what concerns the Worst IFRS stage score and Rating Score migrations, the pattern illustrated by the respective stem plots C.7, reporting the number of entities counted for each type of migration, was rather coherent with the status of entities recorded in April 2023. Indeed, the charts indicated that the majority of the clients did not exceed the pre-defined threshold for both attributes and, therefore, they did not face any negative migration over time. However, it is important to underline that the number of entities that did present a low-quality credit level or a high-risk profile, represented by their IFRS and Internal Rating scores, was slightly more important than Cluster 1.

- **Cluster 3**:

  1. *Default/Watchlist Status*: As of April 2023, the entity in question was not reported to be in a financial distress condition and was not included into the bank’s watchlisting procedures.

  2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: From the investigation of the growth of Allocated Limit, EAD, and Outstanding Amount

over the months, it was inferred that the entity either encountered an extension in the bank-provided credit limit or obtained a new loan agreement involving a higher credit limit (Figure C.9). This expansion was followed by a more modest rise in Outstanding Amount and EAD as well. These increases were not reflected in terms of the RWA and Expected Loss, which reported a significant negative growth, instead.

3. *Average number of monthly triggers and news articles*: The entity in question reported an average number of monthly external triggers equal to 0 and an average of 0.8 monthly internal triggers, implying that the client typically would typically activate at least one trigger almost every month. Moreover, the client was usually referenced in one negative article every month, which suggests that the entity in question could be a well-known organisation (Figure C.9).

4. *Worst IFRS stage and Internal Rating score migrations*: In relation to both the Worst IFRS stage and Internal Rating Scores, there were no instances of negative or positive migrations observed for this borrower. The client had consistently preserved a high credit quality and low-risk profile throughout the study.

5. *Most recent allocated limit and outstanding amount*: Through the analysis of April's allocated limit and outstanding amount Figure C.6 demonstrated that the client had been granted with substantial limits, that were almost fully utilised and consumed. This observation further supports the hypothesis that the entity could potentially represent a prosperous and sizable corporate entity.

- **Cluster 4**:

1. *Default/Watchlist Status*: Although an increased number of the entities presented a critical situation, the majority of the borrowers involved were associated with ordinary statuses. Still the percentage reported in Figure C.9 for this very subset of clients was slightly less significant than Cluster 2.

2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: Much like the findings discovered for Cluster 2, this category also encompassed clients who have undergone relevant changes over time across all five attributes. However, in this case, the changes in question entailed a noteworthy decline in EAD, RWA, Expected Loss, Allocated Limit, and Outstanding Amount. Specifically, the negative growth detected for the EAD, RWA, and Outstanding Amount attributes ranged between -1.8 and -1.3 (Figure C.9). In contrast, the statistics for the Expected Loss and Allocated Limit showed a relatively minor and less marked decrement.

3. *Average number of monthly triggers and news articles*: Concerning the monthly external triggers, it seemed that these borrowers exhibited a limited frequency of trigger occurrences on a monthly basis. As depicted by the feature's corresponding distribution box plot C.6, the average number of monthly external triggers was mostly

zero for the majority of entities, with a maximum of 0.5 for the outliers. Also the data regarding the average number of monthly internal triggers did not indicate a concerning or dangerous scenario. However, the values reported for the attribute's 25^{th} percentile and median were slightly above the overall average between clusters and, specifically, Cluster 2 (Figure C.9). This suggested that Cluster 4 was comprised of more borrowers with a higher likelihood of generating one trigger per month.

4. *Worst IFRS stage and Internal Rating score migrations*: Even though the cluster included instances of entities that experienced both positive and negative migrations, mainly for the IFRS stage, the majority of the clients were actually assigned with "good" credit levels already from their respective starting month.

- **Cluster 5**:

  1. *Default/Watchlist Status*: Once again, a substantial portion of entities were allocated non-alarming activity statuses in April 2023, composing, approximately, 94% of that total (Figure C.9). However, the proportion of clients categorised as in default or watchlisted was greater than Cluster 4.

  2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: The situation that arose from the statistical analysis of growth-related attributes in Figure C.9 resembled the scenario explored in Cluster 1. This set of clients consisted of borrowers who either experienced negative growth or reported no growth at all, with a distinct decrease observed for Expected Loss and RWA. Nevertheless, the values documented for the 25^{th} percentiles of each attribute were notably lower and more marked compared to Cluster 1. This indicated that this segment contained entities that encountered more substantial depreciations.

  3. *Average number of monthly triggers and news articles*: No noticeable distinction was identified in the count of monthly external triggers flagged for this subgroup of borrowers. On the contrary, the data related to the percentiles and the median for the mean number of internal triggers indicated that, within this segment, a greater number of customers scarcely raised any internal triggers on a monthly basis (Figure C.9). This observation was derived by the fact that the recorded values were significantly smaller compared to those reported for the other clusters.

  4. *Worst IFRS stage and Internal Rating score migrations*: Regarding the migration of the clients' Internal Rating scores, all the entities included experienced a positive shift, transitioning from a "bad" score to a "good" one over time. This migration was also evident in the context of Worst IFRS stage; however, it was observed that, in practice, the majority of the entities had been assigned with a Stage 1 of the IFRS standard right from the beginning.

- **Cluster 6**:

1. *Default/Watchlist Status*: As reported by Figure C.9, the percentage of entities labeled as either "In Default" or "Watchlist" in April 2023 was higher than Cluster 5, constituting a 12% of the total.

2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: Through the analysis of the attributes linked to overall growth, it was discerned that a significant portion of the clients exhibited either a decline in growth or remained in a state of stagnation. This trend was notably marked for metrics such as Expected Loss, RWA, and Outstanding Amount, as reported by the medians and 25^{th} percentiles in Figure C.9. Nevertheless, a deeper examination of the 75^{th} percentiles unveiled that a subset of these clients experienced modest yet positive growth patterns in relation to Allocated Limit and Expected Loss.

3. *Average number of monthly triggers and news articles*: In relation to the average number of external and internal triggers, the data concerning the median and other statistical measures defined a notably heightened sense of concern and importance in comparison to the previous segments. Firstly, since the 75^{th} percentile associated with monthly external triggers exceeded the value 1, it was assumed that this particular group comprised a larger proportion of entities that regularly activated one or more external triggers on a monthly basis (Figure C.9). Furthermore, a more detailed examination of the box plots provided in Figure C.6 revealed that the statistics related to the average monthly number of internal triggers were the most elevated among the various clusters. Finally, the count of positive and negative news articles generated on a monthly basis showed a remarkably unusual and outstanding nature, that ranged between 2 and 4 articles for the majority of clients (Figure C.9).

4. *Worst IFRS stage and Internal Rating score migrations*: Overall, although with a lesser extent when compared to the previous cluster, Cluster 6 encompassed entities that were designated IFRS stages and Rating Scores falling below the predetermined thresholds both at the beginning and the ending of the investigation.

- **Cluster 7**:

1. *Default/Watchlist Status*: As indicated by the stem plot in Figure C.7, this segment represented the second cluster with the highest number of distressed or watchlisted clients in April 2023.

2. *Total Allocated Limit, EAD, RWA, Expected Loss and Outstanding Amount growths*: Regarding the EAD, Allocated Limit and Outstanding Amount growths, the main characteristics that could be noticed at a glance concerned the lack of growth or the negligible decreases that these features exposed. On the contrary, the data regarding the descriptive statistics of the Expected Loss and RWA growths, shown in Figure C.9, suggested that the clients included in this specific cluster either remained stable or dealt with a significant monthly increase for these attributes, symbolising a

potential surge of the associated risk exposure.

3. *Average number of monthly triggers and news articles*: The investigation of the average number of external and internal triggers raised for these entities did not highlight any relevant information about the risk involved with these clients, as the statistics were not particularly informative or insightful compared to the other clusters.

4. *Worst IFRS stage and Internal Rating score migrations*: Fueling the theory that the customers involved posed inherent risk were the negative migrations that were identified for the IFRS and the Internal Rating standards. Specifically, every client assigned to Cluster 7 transitioned from a favorable to an unfavorable Internal Rating score (Figure C.7). This shift also manifested in terms of Worst IFRS stage, although a minor portion of entities still preserved their initial positive credit quality levels.

5.5.3. HIGHLIGHTS FROM CLUSTERS' DESCRIPTIVE STATISTICS ANALYSIS

The descriptive statistics analysis of each cluster offers a comprehensive view of the unique trends and peculiarities inherent to different segments within the dataset. From this in-depth and detailed investigation, several important aspects and insights could be derived. Firstly, the analysis revealed that the entities assigned to each segment were indeed characterised by similar traits and properties. However, as the traits in question were only related to a limited subset of the features involved, these common patterns did not define an all-around similarity within each cluster. In fact, it was observed that, for most segments, the reported statistics appeared to be consistent and less sparse only for a certain number of attributes, representing the most significant and meaningful features of the respective cluster. For this reason, it was not always possible to properly categorise and label specific groups of borrowers merely by their shared characteristics and attributes.

Nevertheless, despite this limitation, it was discovered that, in reality, several clusters did exhibit more coherent and concordant scenarios for most of the variables, thus facilitating their interpretation and classification. For instance, based on the description of Cluster 3 (Section 5.5.2), generated from the implementation of K-Means over the so-called uncorrelated dataset, it was discerned that the majority of the entities involved did present a more favourable and promising profile compared to the other clusters. Indeed, the statistics of Figure C.4 revealed that the clients in question were assigned to regular and not critical status (except for a small and negligible percentage) and were most likely dealing with a depreciation of their PD or LGD, despite engaging with a new loan. Moreover, it was also noticed that Cluster 3's clients would not generate any trigger on a monthly basis and presented a positive credit quality throughout the whole study.

On the contrary, for the elements composing Cluster 2, the exploration of the respective descriptive data unveiled a more concerning and problematic picture. By including over 30% of distressed or watchlisted clients, generating more than 1 internal warning every month and

presenting high-risk credit ratings and scores both in their starting and ending months, Cluster 2 was indeed defined as the most unfavourable and adverse segment among the ones created.

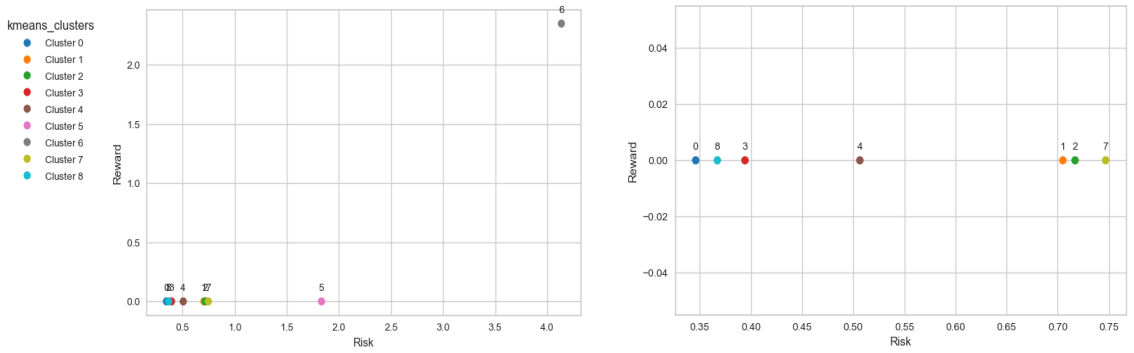
5.5.4. RISK-REWARD ANALYSIS

Since the ARIA tool was developed with the purpose of enhancing the supervision of customers at risk and detecting potential credit incidents at an early stage, the implementation of a secondary analysis focused entirely on the aspect of early warning triggers was deemed necessary. The objective of this investigation was to obtain a deeper understanding of the financial health and behaviour characterising each cluster through a risk-reward analysis that focused on the average number of positive and negative triggers raised each month. As detailed in Chapter 4, the risk component of this analysis was represented by the average of monthly counted negative triggers raised within the cluster, which encompassed negative news articles and internal/external triggers. Conversely, the reward component was associated to the average number of positive articles identified for each segment. The comparison between these two metrics enabled the researcher to evaluate the balance between risk exposure and business potential involved when engaging with a particular subgroup of clients. However, it was noted that, due to the broader intra-clusters distribution and inter-cluster overlap that these features manifested for numerous segments, these attributes had a limited influence on segmentation of the involved individuals. Moreover, it is important to underline that the aforementioned calculations were exclusively computed for entities falling within the interquartile range of each attribute. Therefore, these two fundamental approximations must be taken into considerations when interpreting the results obtained from this analysis, as they represented pivotal drivers that deeply determined the analysis trustworthiness.

UNCORRELATED DATASET: K-MEANS

From the analysis of Figure 5.13a, it was discovered that one single cluster was composed by entities featuring an average monthly count of positive news articles greater than 1, referred to as Cluster 6. However, this very cluster also manifested a notably elevated occurrence of negative triggers on a monthly basis compared to the other clusters, thereby rendering it notably risky and unreliable.

A different behavior was exhibited by the remaining clusters. Indeed, it appeared that the segments in question did not generate any positive article warning and displayed low averages of monthly negative triggers as well. Among them, Cluster 5 was the only exception, showing an average number of monthly negative triggers approximately equivalent to 2. In addition, Figure 5.13b indicated that Cluster 1, Cluster 2, Cluster 7 and, to a somewhat lesser extent, Cluster 4, generated approximately one trigger almost every month. As for the other segments, the chances of flagging either a trigger or a negative article were notably less critical, especially for Cluster 0.



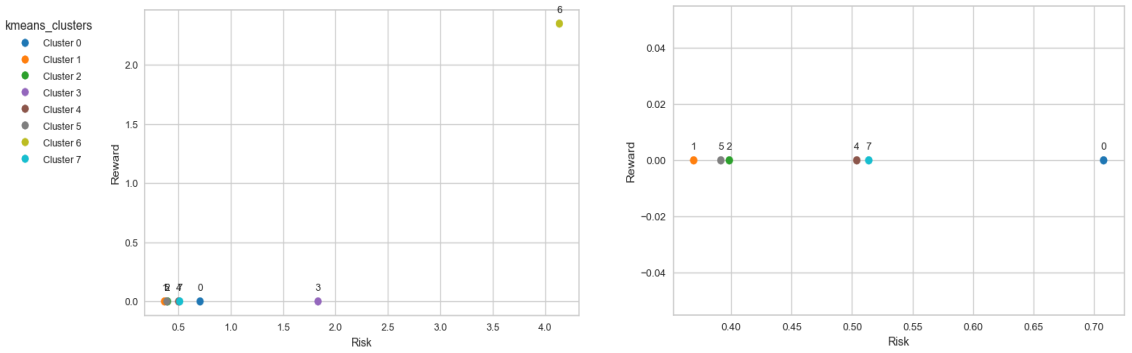
(a) Overview

(b) Zoom in of overview

Figure 5.13: Risk-reward analysis for clusters generated from the implementation of K-Means on the dataset obtained from correlation analysis

PCA DATASET: K-MEANS

The scenario depicted in Figure 5.14a, which showcases the outcomes derived from the application of K-Means to the PCA-transformed dataset, strongly resembled the situation analysed for the uncorrelated dataset in the previous subsection. This observation suggested that minimal and negligible variations occurred when utilising these two datasets separately. In fact, once again, only one single cluster, i.e. Cluster 6, showed positive values for the reward attribute, while, at the same time, recording a great number of negative triggers each month.
Moreover, similar to the findings discovered in the previous exploration for Cluster 5, an analogous behavior and pattern was detected for Cluster 3, as approximately 2 negative triggers were also raised for this group of entities every month.
Regarding the other formed segments, Cluster 0, with a rate of over 0.7 negative triggers each month, displayed the highest probability of triggering either one alert or a negative article on a monthly basis. The values assigned to the risk attribute for the remaining subsets of clients where relatively less critical and concerning. However, Cluster 4 and Cluster 7 still showed an average number of negative triggers exceeding 0.5. Finally, the segment that was characterised by the lowest scores among all the clusters generated, indicating a more financially sound and stable profile, was indeed Cluster 1.



(a) Overview

(b) Zoom in of overview

Figure 5.14: Risk-reward analysis for clusters generated from the implementation of K-Means on the PCA-transformed dataset

5.5.5. RISK EXPOSURE ANALYSIS

The focus of this exploration remained on examining the outcomes generated by the K-Means clustering algorithm for each dataset utilised as a foundation of this research study. Moreover, in the case of the PCA-transformed dataset, the findings discovered were also compared to the average Total Outstanding Amount that every clustering presented in April 2023, in order to obtain a deeper knowledge of the risk involved within each subset of entities.

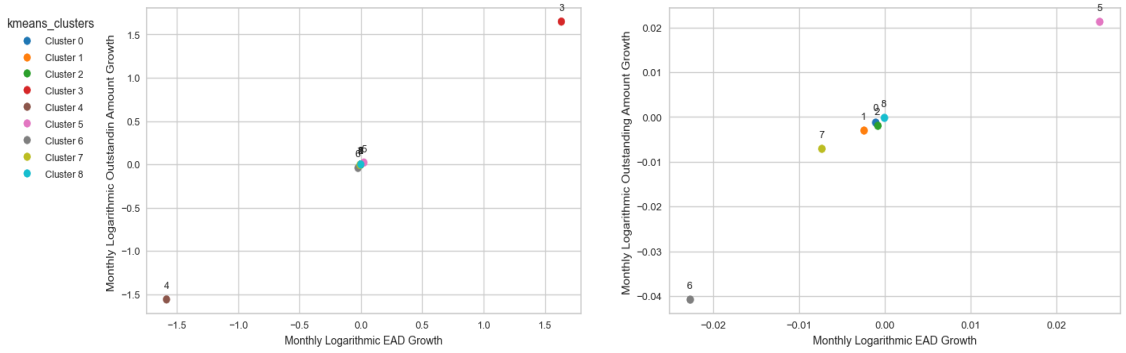
Likewise in the Risk-Reward analysis, it is important to emphasise that the results and insights drawn from this exploration should be viewed with caution, as the features in question exhibited a limited significance and contribution for most of the clusters. As a result, the reliability and trustworthiness of the conclusions drawn from this analysis may be limited.

UNCORRELATED DATASET: K-MEANS

Figure 5.15a aims at shedding light to the average risk exposure involved within each cluster by illustrating the average Outstanding Amount growth in function of the respective EAD growth computed for each segment. By analysing this very image, it was immediately noticed that two distinct clusters, namely Cluster 3 and Cluster 4, exhibited more significant variations in logarithmic fashion in relation to both attributes. More specifically, Cluster 4 was marked by substantial decreases, whereas Cluster 3 showcased noteworthy increases instead. This observations were easily discernible given also that the data points corresponding to these groups of entities were significantly distant from the remaining indicated segments. Moreover, despite their divergent growth trends, the reported ratios between the EAD growth feature and the Outstanding Amount growth for both clusters showed an overall similarity and were approximately equal to 1, hinting that no increment of risk exposure occurred for these subsets of borrowers. The reason for these atypical logarithmic increases and decreases is rooted in the way the data was processed and manipulated when customers indicated a 0 value in a feature either at the beginning or end of their activity period.

For what concerned the remaining clusters, as observed in Figure 5.15b, it was noticed that the majority of them displayed comparatively less pronounced declines in both attributes and

that these reductions were often aligned in terms of magnitude, with the exception of Cluster 8, Cluster 2 and Cluster 6. As reported in Table 5.2, the former cluster was characterised by small negative decreases for both attributes. However, the cluster's positive EAD-Outstanding Amount ratio appeared to be particularly marked, symbolising a divergence between the changes in credit balances and the associated potential risk exposure. Cluster 6, instead, demonstrated a more noticeable decrease for both features and an EAD-Outstanding Amount ratio below 1, entailing that, when compared to its corresponding Outstanding Amount decrease, a more modest negative variation in the EAD was detected.
Finally, beside Cluster 3, the only segment that was characterised by positive growths for both features was indeed Cluster 5, as illustrated by Figure 5.15b. Not only these entities presented relevant increments in terms of EAD and Outstanding Amount, but also a potential increase of their associated risk exposure as well. This conclusion was derived by the examination of the EAD-Outstanding Amount ratio which appeared to be notably greater than 1 for this very segment.



(a) Overview

(b) Zoom in of overview

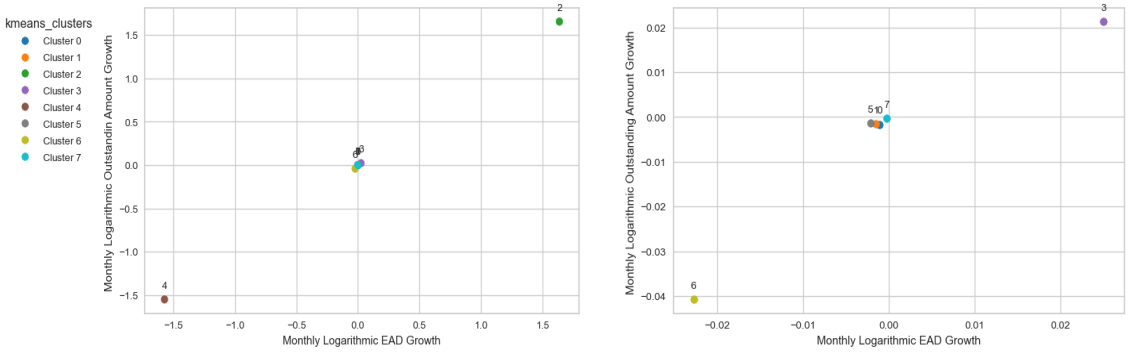
Figure 5.15: Risk exposure analysis for clusters generated from the implementation of K-Means on the dataset obtained from correlation analysis

Table 5.2: Summary of the K-Means clusters' average Outstanding growth, EAD growth and EAD-Outstanding ration for the uncorrelated dataset

K-Means Cluster	**EAD Growth**	**Outstanding Growth**	**EAD-Outstanding Ratio**
0	-0.001100	-0.001088	1.01
1	-0.002487	-0.002960	0.84
2	-0.000866	-0.001900	0.45
3	1.633508	1.648977	0.99
4	-1.583113	-1.558997	1.01
5	0.025014	0.021329	1.17
6	-0.022743	-0.040768	0.56
7	-0.007402	-0.007041	1.05
8	-0.000119	-0.000052	2.29

PCA DATASET: K-MEANS

Once again, the scenario portrayed in the clusters formed by applying K-Means to the dataset derived from PCA implementation closely resembled the one examined in the previous subsection. As depicted in Figure 5.16a, there were two primary clusters that displayed more noticeable average increases and decreases for both attributes. These changes, however, were also reasonably congruent with each other. A more focused view of the clusters distribution, illustrated in Figure 5.16b, highlighted segments with more regular growth patterns and mainly characterised by milder negative logarithmic growths. These decreases, however, were proportionally aligned only for Cluster 1 and Cluster 7, as outlined in Table 5.3. Conversely, the other subsets of entities showed a more substantial negative growth for either one of the two attributes. Indeed, Cluster 0 and Cluster 6 reported a greater decrement in terms of the Outstanding Amount, while Cluster 5 demonstrated a considerably critical reduction in the EAD in comparison to its corresponding Outstanding Amount growth. Regarding the former two cluster aforementioned, Table 5.3 demonstrated that, not only Cluster 6 presented a more significant decline in the Outstanding Amount compared to its corresponding EAD growth, confirmed by the smaller value reported for the EAD-Outstanding Amount ratio, but also it appeared that the entities that composed this specific were characterised by more significant Total Outstanding Amounts in April 2023 as well. Finally, a concerning scenario emerged when examining the EAD-Outstanding Amount growth ratio for Cluster 3. Indeed, it was found that this particular group of elements displayed significant increases in both attributes, with a notably higher positive growth in the EAD, indicating an amplified level of the associated risk exposure. In addition, the Total Outstanding Amount recorded for this very subsets of borrowers represented the most critical outstanding balance across all clusters.



(a) Overview

(b) Zoom in of overview

Figure 5.16: Risk exposure analysis for clusters generated from the implementation of K-Means on the PCA-transformed dataset

Table 5.3: Summary of the K-Means clusters' average Outstanding growth, EAD growth, EAD-Outstanding ratio and Outstanding Amount for the PCA-transformed dataset

K-Means Cluster	**EAD Growth**	**Outstanding Growth**	**Total Outstanding Amount**	**EAD-Outstanding Ratio**
0	-0.001111	-0.001643	6.371868e+04	0.68
1	-0.001499	-0.001511	1.801486e+06	0.99
2	1.638770	1.654103	2.912977e+06	0.99
3	0.025014	0.021329	7.071330e+10	1.17
4	-1.573920	-1.548339	0.000000e+00	1.01
5	-0.002150	-0.001378	3.174296e+04	1.56
6	-0.022743	-0.040768	2.231486e+07	0.56
7	-0.000252	-0.000238	3.341819e+04	1.06

5.6. SHAP ANALYSIS

The values illustrated in Figure 5.17 report the SHAP values indicating each feature's contribution on the clustering model for a specific cluster. The values were computed on two different trials related to distinct random samples of the original data. Furthermore, the values in question were calculated using the dataset obtained from the correlation analysis procedure, and represented the most significant attributes that characterised Cluster 4. From the examination of the multiple outcomes of the Kernel SHAP method, regarding different subsets of the initial dataset, like the ones depicted in the charts of Figure 5.17, it was discovered that the results produced differed from each other only in a marginal way and that each feature maintained its importance in almost every attempt. This phenomenon, however, was more evident for the most significant attributes and more visible for the clusters that presented a higher density of entities, since the calculation of the SHAP values was based on a larger subgroup of clients. For instance, if considering the values generated for Cluster 4, it was observed that, on different trials, less impactful attributes, like the "rating_bad_good_migration" feature, presented different SHAP values. This discrepancy was most likely caused by the undersampling process of the data, which allowed the computation of the SHAP values only for a smaller subset of the entities, and, therefore, included different distributions of features' values on each trial.

Despite this limitation, it was discovered that the results obtained from the SHAP analysis were mostly consistent with the observations drawn from the statistical data exploration. In fact, in the case of Cluster 4, it appeared that low scores and mild values for the growth-related features, i.e. the EAD growth, Outstanding Amount growth, the RWA growth and the Expected Loss growth, played a pivotal role and represented the main driving factors for the clustering. On the contrary, the data concerning the average number of monthly triggers and news articles or the credit limit and activity status assigned in April 2023 contributed to the segmentation with a minor influence. Likewise, Figure 5.18 portrays the summary plot of the features contribution concerning, respectively, Cluster 2, computed for two different samples randomly selected from the initial dataset. Once again, the most influencial and decisive attributes, characterised by higher SHAP scores, remained coherent with each other across the different tri-

als and preserved their permutation importance. Moreover, for this specific cluster, it was observed that high values of the “ifrs_bad_bad_migration", “rating_bad_good_migration" and “default_watchlist_status" features and low values of the “ifrs_good_good_migration" and “rating_good_good_migration" were deemed to exhibit a greater contribution to the clustering of the entities involved. This could be derived from a more marked degree of variability across the data points that these features presented. As a consequence, these attributes might have been able to differentiate between data points more effectively, by exhibiting distinct patterns or more defined ranges of values. On the contrary, similarly to what emerged for the previous cluster, attributes concerning the average number of monthly triggers or the total allocated limit recorded in April 2023 seemed to have had the least significance instead.
More results of the SHAP analysis, conducted for different clusters, are provided in the Appendix D.

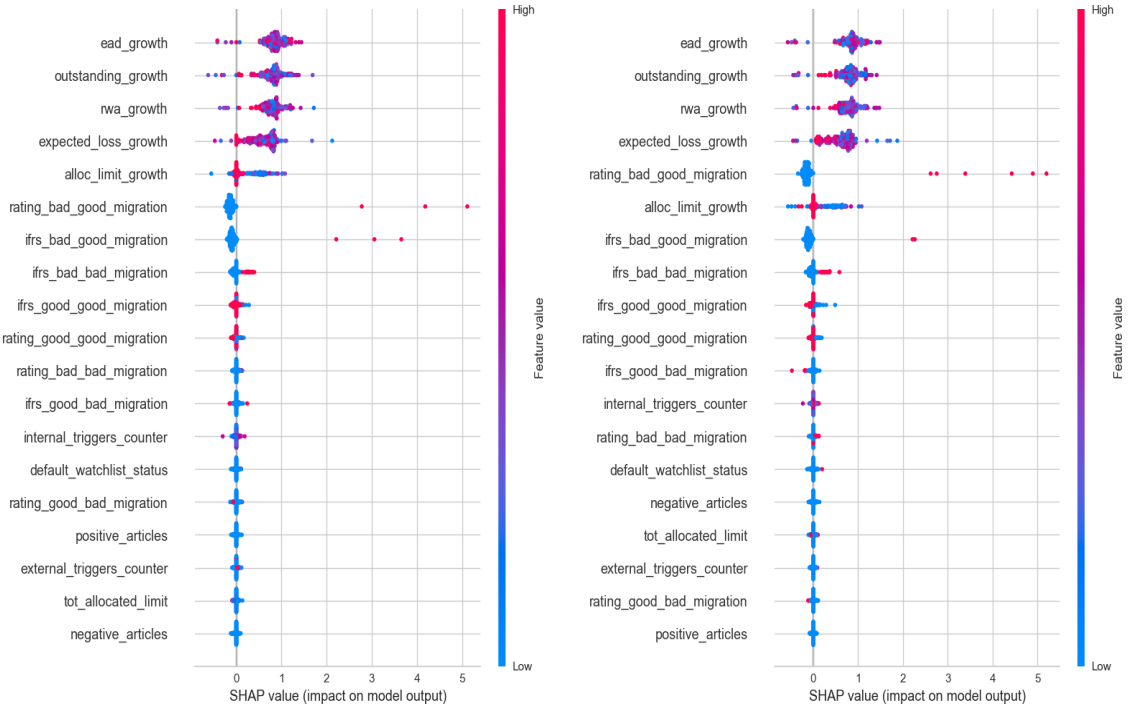


Figure 5.17: SHAP values of the most significant features for Cluster 4 generated with K-Means and the dataset obtained from correlation analysis

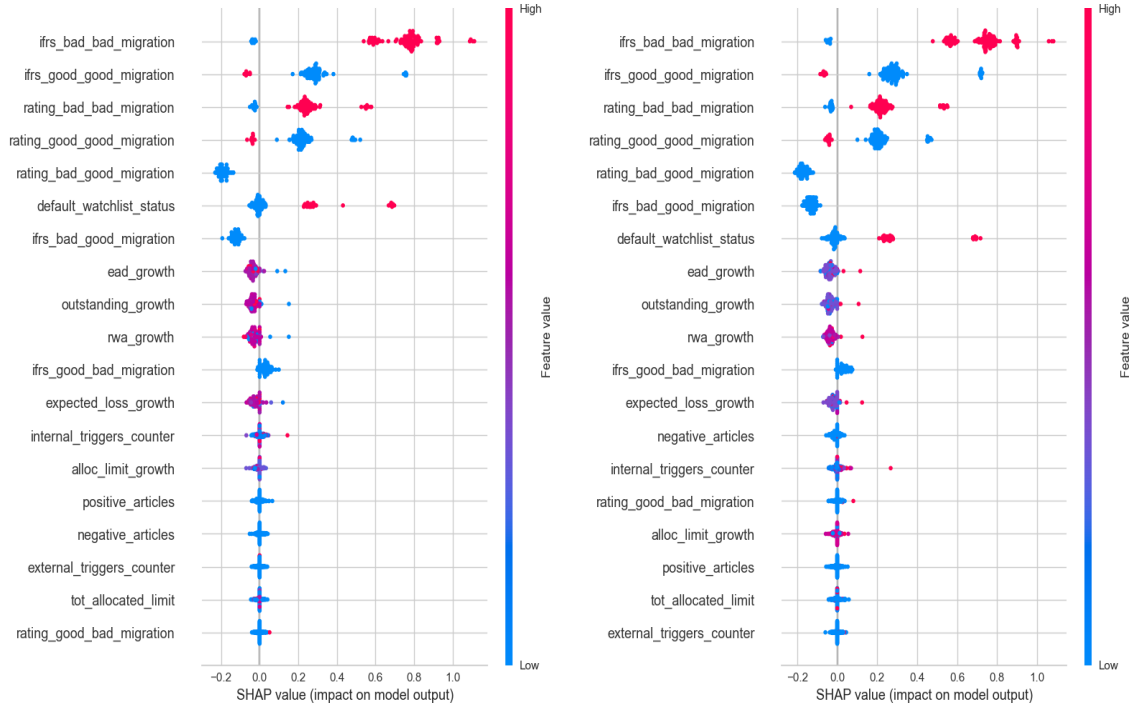


Figure 5.18: SHAP values of the most significant features for Cluster 2 generated with K-Means and the dataset obtained from correlation analysis

5.7. MODELS VALIDATION RESULTS

This final subsection aims at providing an overview and brief summary of all the validation techniques that have been deployed to assess and determine the quality and reliability of the analytical methods used.

1. **K-Means Number of Clusters Validation**: as the effectiveness and efficiency of the K-Means clustering algorithm heavily relies on the initial selection of the most appropriate number of clusters, several methodologies have been developed to support this decision and to aid researchers and analysts in making informed choices. The KElbowVisualizer, offered by Python's Yellowbrick library [3], represents a valuable validation tool that automates this process by analysing the relationship between different numbers of clusters and the respective intra-cluster inertia obtained. Based on the outcomes of this analysis, the tool is also capable of identifying the so-called elbow point, representing the optimal number of segments. From the implementation of the KElbowVisualizer, the number of subgroups selected for the uncorrelated dataset and the PCA-transformed dataset were, respectively, equivalent to 9 and 8 clusters.

2. **DBSCAN Hyperparameters Validation**: for what concerned the application of the DBSCAN model, several rules of thumb have been deployed for the determination of the optimal values of both ϵ and MinPoints, the main hyperparameters of this specific clus-

[3] https://www.scikit-yb.org

tering model (Section 3.3.3). For instance, one approach involved setting MinPoints to be twice the dimension of the dataset and plotting the distances of each data point to its k-nearest neighbor in ascending order, where k represented the initial MinPoints value chosen. The point on the plot with the most pronounced curvature was then considered as the optimal value for ϵ. However, since each one of these techniques resulted in generating only 2 or 3 segments, it was believed that none of the methods in question was indeed suitable for the type of dataset under examination. Therefore, a novel and custom approach was adopted, focused on the analysis of the trade-off between the number of clusters and the respective Silhouette Score obtained for different combinations of these two hyperparameters. Based on the results observed from this investigation and the researcher's qualitative evaluation, the values selected for ϵ and MinPoints were equivalent to 5.0 and 3 for the uncorrelated dataset and 4.0 and 8 for the PCA dataset.

3. **Clusters Quality Evaluation**: after the application of the two clustering algorithms for each one of the datasets obtained from the dimensionality reduction procedures, an assessment of the clusters quality was deemed necessary in order to validate the actual performance of the algorithms adopted. This was achieved by computing three different metrics, namely the Silhouette Score, the Davies-Bouldin Score and the Calinski-Harabsz Score, capable of quantifying the overall intra-cluster compactness and the inter-cluster separation of the segments formed (Section 4.5.1). From this analysis, it was discovered that K-Means outperformed DBSCAN across all metrics, especially in the case of the PCA-transformed dataset, indicating the creation of tighter and more distinct clusters.

4. **SHAP Analysis Validation**: finally, the last study of the whole research related to the exploration of each feature's contribution and impact on the K-Means clustering process. The analysis was conducted with the help of Python's SHAP library [4], capable of calculating the features' SHAP values and defining their influence on the model. Although the library was not specifically built for unsupervised learning models, the use of the model-agnostic Kernel SHAP explainer allowed the research to overcome this limitation. Nevertheless, because of the demanding computational resource and complexity that the library requires for computing the SHAP values for each variable and each cluster, an alternative approach was implemented. The method in question consisted in fitting the explainer model on multiple random subsets of the initial dataset, containing 1,000 samples, and comparing the results generated at each trial in order to validate the integrity and consistency of the outcomes obtained.

[4] https://shap.readthedocs.io

6

CONCLUSION

6.1. LESSON LEARNED

In conclusion, this research consisted in an investigation on the integration of EWS into a new CS model for WB clients, in order to identify a potential segment of customers that presents the ideal profile for potential business up-selling opportunities and credit limit extensions. The main research question, representing the foundation of the whole study, is answered based on the acknowledgements derived from the following subquestions:

> *How can the information obtained from the EWS be processed in order to be integrated in the CS model?*

Firstly, in order to align the initial business objectives of the research to the outcomes of the model developed, a number of crucial requirements have been defined, delineating the main properties that the artifact should present and the principles it should adhere to. The desired properties outlined consisted in the following criteria: the CS model should be risk-oriented and generate clusters of entities sharing similar risk levels, the segments created must be distinguishable and recognisable, and, finally, the insights derived from the exploration of the clusters formed should allow the stakeholders involved to develop actionable strategies and facilitate their decision-making processes.
On the basis of these requirements, a number of features have been designed and engineered in order to obtain a complete understanding of the risk scenario displayed by every client. The variables in question aimed at determining the financial health of each borrower from different perspectives: the average number of triggers raised on a monthly basis, the growth detected for multiple indicators related to the client's risk exposure, the evolution of the entity's credit risk status and, finally, the most recent financial activity status, outstanding balance and allocated credit limit of the borrower.

> *How can customer segments be generated in an automated manner?*

Secondly, based on the insights drawn from a systematic review of the literature related to the application of both CS and EWS models, two different clustering algorithms have been deployed in order to determine the most appropriate and efficient model for the set of data used. The algorithms selected were, respectively, K-Means and DBSCAN. Moreover, two separate dimensionality reduction techniques, namely the Correlation Analysis and PCA, have been deployed in a parallel with the aim of identifying the approach that was deemed to be the most effective and insightful.
Although the validation of the results obtained can be rather challenging in the absence of explicit target labels, several validation techniques have been utilised to quantify the performance of each model for different values of their hyperparameters. In the case of K-Means, the implementation of the so-called Elbow Method enabled the selection of the most appropriate number of clusters based on the analysis of the variation of clusters compactness. For what concerned the DBSCAN model, it was observed that none of the conventional rules of thumb for the selection of the optimal hyperparameters values were indeed applicable for the dataset available. Therefore, a different approach was adopted, focused on the identification of the most suitable trade-off between the number of clusters generated and the corresponding Silhouette Score detected.
Finally, the computation of the SHAP values for multiple random subsets of the dataset and the different clusters formed proved to be a meaningful method that provided relevant insights on the significance and the contribution that each feature had on the clustering process.

> *How can the quality of the clusters generated be assessed and how can the segments obtained be interpreted?*

The evaluation of the clusters generated and the performance of the models adopted was achieved through the combination of a total of five different analyses. First, the quality of the segments obtained was assessed by measuring three different metrics defining the compactness and the separation of the groups, namely the Silhouette Score, the Davies-Bouldin Score and Calinski-Harabasz Score. Then, an investigation of the clusters densities and a statistical analysis of the clients' distribution within each cluster and for every feature was implemented for both the PCA dataset and the uncorrelated dataset. According to these experiments, it was observed that K-Means significantly outperformed the DBSCAN clustering model by producing significantly more cohesive and separated clusters, hinting a stronger differentiation and definition between the segments. Moreover, the implementation of a PCA-transformed dataset slightly enhanced the results as well. From these analyses, it was also discovered that the outcomes of the deployment of the DBSCAN algorithm appeared to be rather inconclusive and inaccurate, as no real and visible distinction could be detected across the various segments.
Next, a risk-reward analysis based on, respectively, the average number of monthly negative triggers and positive articles raised by each segment allowed to gain a better understanding of the risk associated with every subset of entities created. Finally, from an exploration and comparison of the cluster's average EAD and Outstanding Amount growths, the increased or decreased risk exposure that the entities involved exhibited was eventually assessed.

Given the results obtained from the development of this new and innovative approach for CS, it is important to underline that, due to the very nature of the clustering algorithms and the underlying structure of the data used, none of the clusters generated from the implementation of this model presented a favourable and consistent scenario for every single feature engineered. Specifically, from the analysis of the features' importance implemented for all the segments created, it was discovered that only a limited number of attributes presented a significant contribution to the segmentation of the entities included each time. As a consequence, this phenomenon led to the formation of clusters characterised by entities sharing common traits and behaviours for only a minor subset of attributes. Because of this limitation, the integration of managers' and stakeholders' business domain knowledge represents a valuable and essential asset that would enable the prioritisation of the most relevant features and define the appropriate win-loss trade-off that could lead to the identification of the most profitable and high-value segment of borrowers.
Nevertheless, from the in-depth exploration of the clusters obtained using the K-means model, it was still observed that specific subgroups of clients presented a more promising and positive scenario in comparison to the other segments generated. Certainly, clusters like Cluster 3 (Section 5.5.2) were perceived as potential focal points for future business prospects as they primarily consisted of entities with regular status, limited number of monthly triggers, reassuring credit risk ratings and stagnant risk exposure.

6.2. PRACTICAL AND SCIENTIFIC CONTRIBUTIONS

From a practical perspective, the main contribution that this research was able to achieve relates to the improved and deepened understanding of the financial health of ING's WB client base. Indeed, by introducing the information regarding the number of triggers and warnings that each client generated, the study managed to discover relevant insights regarding the credit risk situation of specific subsets of similar clients and shed light to underlying complex customer dynamics. As a result, the information derived from these acknowledgements can equip decision-makers with the capability of re-adapting their strategies, tailor their offerings and monitor their credit portfolios as a response to the scenario exhibited by each segment. Moreover, the findings of this investigation also contribute to assessing the actual efficacy of ARIA's EWS by unveiling all potential cause-and-effect relationships among different attributes and highlight undetected high-risk situations.
In addition, an application of EWS for CS purposes similar to the one developed within this study, may not only enhance the risk management practices of banks' commercial clients, but, if properly adapted to the settings of the implementation, with appropriate and relevant features, it can also lead to significant contributions that extend well beyond the realm of WB. For instance, through the implementation of EWS, healthcare providers would be capable of monitoring and controlling the health deterioration of their high-risk patients. As a consequence, the segmentation of clients based on the number of warnings generated over a specific period of time would allow hospitals and health institutions to detect patients with aggravated health

conditions and, therefore, improving patient care while limiting the risk of medical emergencies. In the case of E-commerce and retailers, instead, by clustering entities on the basis of customers' deteriorating product demand and inactivity signals, stakeholders would be provided with a more in-depth understanding of potential churning risks within the client base. Subsequently, this knowledge would help retailers offer more personalised marketing promotions to specific segments and increase customer loyalty. With regards to the energy and utilities sector, early warning signals could serve as detectors of high energy consumption and of potential failures or malfunctions of the equipment provided. Then, the information obtained from the system could be leveraged by energy distributors to power CS applications and to initialise energy conservation and efficiency campaigns based on the communication of the energy-saving measures and appropriate energy usage to the respective subgroups of customers identified.

Finally, this study contributes to the existing body of literature by bridging the gap between EWS and CS practices. Indeed, the integration of EWS into CS introduces an innovative and advanced perspective related to strategic risk control. Traditionally, CS focuses on the categorisation of clients based on historical and behavioural data. However, with the incorporation of warning signals, a new dimension of deepened risk understanding and monitoring is introduced in the field of segmentation. Moreover, this significant advancement opens new discussions for research and discourse on different aspects. For instance, scholars can delve into the investigation of the prediction of more robust and efficient warning indicators or analyse the ethical matters sorrounding the use of such data for categorising clients. In summary, this approach is not only capable of enhancing the practical utility of segmentation procedures but also stimulates intellectual research on leveraging EWS for marketing and customer management purposes.

6.3. LIMITATIONS AND FUTURE RESEARCH RECOMMENDATION

Despite the valuable insights garnered by this study, it is important to acknowledge a number of limitations that have deeply impacted the outcomes achieved and that should be taken into consideration when interpreting the study's findings.

1. **Utilisation of one citation database, specific search filters and qualitative analysis to review the literature**: Firstly, the scope of the systematic literature review was limited by the utilisation of one unique citation database, Scopus, to collect relevant articles. In addition, the application of specific search criteria and the qualitative selection of appropriate articles, solely based upon the researchers' personal perceptions and evaluations, may have excluded insightful studies from the current investigation.

2. **Literature limitations**: The systematic review suggested that the existing literature concerning both EWS and CS practices is still to be considered rather restricted and limited. Indeed, it was discovered that these technologies have only recently begun to gain attention and traction from the experts.

3. **Application of two clustering algorithms**: Although the chosen algorithms offer valuable

insights related to the entities presenting similar profiles and characteristics, they might not be able to properly encompass all the common patterns within the dataset. As each algorithm presents its own rationale and criteria for partitioning the data, it could occur that some meaningful segments, that a different method would, indeed, be able to capture, are excluded from the outcomes. Therefore, relying solely on two specific clustering models might restrict the effectiveness and richness of all the potential clusters that genuinely reside within the data. Future research could focus on the application of a diverse set of methods which could potentially reveal additional scenarios and provide a more comprehensive understanding of client financial health situation.

4. **Data quality**: The performance and success of every customer clustering model heavily depends on the quality and accuracy of the underlying data. For this reason, potential inaccuracies and errors within the dataset implemented may have lead to a misclassification of the customers involved or highlighted data inconsistencies that comprised the reliability of the segments generated.

5. **Data preprocessing and manipulation challenges**: Since ARIA is founded on the deployment of three different pipelines that integrate the information needed to generate early warning signals from a number of different sources, intensive and intricate preprocessing efforts were demanded before the actual clustering implementation. Because of the impacts that these important data manipulations had on the final input of the clustering algorithms, it is believed that the choice of preprocessing methods and parameters may have influenced the resulting clusters and potentially lead to suboptimal segmentations. Therefore, future research could invest in alternative data imputation techniques or explore new feature engineering approaches that could enhance the models' accuracy and identify more relevant attributes that could contribute to the generation of more distinct and insightful clusters.

6. **Limited data accessibility**: Finally, the last significant limitation stemmed from the limited accessibility of clients data due to confidentiality reasons. While the quality of the outcomes of a segmentation project heavily rely on the level of comprehensiveness of the dataset implemented, in scenarios where sensitive information could be unveiled, the bank is in charge of limiting clients' exposure and preserving their privacy by restricting the accessibility of entities' private information to both internal and external employees. Nevertheless, these protection procedures can comprise the ability to capture different facets of clients' profiles and characteristics. As a consequence, the shortage of customers' features might have compromised the identification of relevant elements behind specific risk behaviours and may have hindered the clustering model's ability to uncover meaningful patterns, thus undermining its effectiveness.

Regardless of the limitations encountered throughout the research, it is possible to provide some additional recommendations that the ARIA team can take into consideration to further expand the added value of the CS model. Firstly, as already addressed by the last limitation

point, leveraging the data collected from more different and diverse datasets could enrich the information available and deliver more informative segments. For instance, by including market and macroeconomic variables or clients' financial accounting data, a more complete and broad representation of customers' financial health and overall profitability could be achieved.
A second recommendation that this study wants to provide relates to the integration of managers' domain knowledge and expertise for the prioritisation of specific features or the assignment of a pre-defined weight to each attribute designed. This innovative solution could be either introduced as additional input information, in order to ensure that the most important variables can influence the clustering process in a more significant manner, or as in a number of criteria that, if properly aligned with the initial business objectives, can be used to rank the clusters generated and automate the identification of prospective segments of clients.
Finally, future studies could aim at deepen the understanding of the outcomes and rationale of the proposed CS model by adopting a more robust approach for the exploration of the features' contribution throughout the clustering process. Since SHAP libraries are typically designed for supervised models, one alternative and renowned technique for the computation of the SHAP values in the case of clustering algorithms involves the integration of a classifier. Indeed, by training this surrogate model on the underlying dataset, using the cluster labels as the final target variable, the effects of the attributes on the prediction of each cluster label can be visualised and analysed using a number of different plots offered by the tool.

REFERENCES

[1] K. Peffers, T. Tuunanen, M. A. Rothenberger, S. Chatterjee. A design science research methodology for information systems research. Journal of Management Information Systems 24 (2007) 45–78. doi:https://doi.org/10.2753/MIS0742-1222240302.

[2] E. Kristoffersen, O. O. Aremu, F. Blomsma, P. Mikalef, J. Li. Exploring the relationship between data science and circular economy: An enhanced crisp-dm process model (2019) 177–189. doi:10.1007/978-3-030-29374-1_15.

[3] J. Ram, C. Zhang, A. Koronios. The implications of big data analytics on business intelligence: A qualitative study in china. Procedia Computer Science 87 (2016) 221–226. doi:https://doi.org/10.1016/j.procs.2016.05.152.

[4] S. . S. V. . C. T. Kansal, Tushar Bahuguna. Customer segmentation using k-means clustering (2018) 135–139. doi:10.1109/CTEMS.2018.8769171.

[5] O. Raiter. Segmentation of bank consumers for artificial intelligence marketing. International Journal of Contemporary Financial Issues 1 (2021) 39–54. doi:http://dx.doi.org/10.17613/q0h8-m266.

[6] S. Mousaeirad. Intelligent vector-based customer segmentation in the banking industry. CoRR abs/2012.11876 (2020). arXiv:2012.11876.

[7] V. Mihova, V. Pavlov. A customer segmentation approach in commercial banks 2025 (2018) 030003. doi:https://doi.org/10.1063/1.5064881.

[8] Open Risk Manual, Early warning indicators for credit risk, 2023. URL: https://www.openriskmanual.org/wiki/Early_Warning_Indicators_for_Credit_Risk, accessed: February 2023.

[9] I. Klopotan, J. Zoroja, M. Meško. Early warning system in business, finance, and economics: Bibliometric and topic analysis. International Journal of Engineering Business Management 10 (2018) 1847979018797013. doi:https://doi.org/10.1177/1847979018797013.

[10] Britannica, ING Group NV, 2023. URL: https://www.britannica.com/topic/ING-Group-NV, accessed: February 2023.

[11] ING, ING at a glance, 2023. URL: https://www.ing.com/About-us/Profile/ING-at-a-glance.htm, accessed: February 2023.

[12] Deloitte Center for Financial Services and the Center for the Edge, Patterns of disruption: Impact on wholesale banking, 2016. URL: https://www2.deloitte.com/us/en/pages/financial-services/articles/collection-disruptive-strategy-patterns-banking.html, accessed: March 2023.

[13] Ç. Gönül, Ö. Özde. The impact of covid-19 pandemic on bank lending around the world. Journal of Banking Finance 133 (2021) 106207. doi:https://doi.org/10.1016/j.jbankfin.2021.106207.

[14] S. Kumar, A. K. Kar, P. V. Ilavarasan. Applications of text mining in services management: A systematic literature review. International Journal of Information Management Data Insights 1 (2021) 100008. doi:https://doi.org/10.1016/j.jjimei.2021.100008.

[15] S. Moradi, F. Mokhatab Rafiei. A dynamic credit risk assessment model with data mining techniques: evidence from iranian banks. Financial Innovation 5 (2019) 1–27. doi:https://doi.org/10.1186/s40854-019-0121-9.

[16] K. Yuan, G. Chi, Y. Zhou, H. Yin. A novel two-stage hybrid default prediction model with k-means clustering and support vector domain description. Research in International Business and Finance 59 (2022) 101536. doi:https://doi.org/10.1016/j.ribaf.2021.101536.

[17] A. Kaminskyi, M. Nehrey, V. Babenko, G. Zimon. Model of optimizing correspondence risk-return marketing for short-term lending. Journal of Risk and Financial Management 15 (2022). doi:https://doi.org/10.3390/jrfm15120583.

[18] P. K. Jadwal, S. Jain, S. Pathak, B. Agarwal. Improved resampling algorithm through a modified oversampling approach based on spectral clustering and smote. Microsystem Technologies (2022) 2669–2677. doi:https://doi.org/10.1007/s00542-022-05287-8.

[19] M. R. Machado, S. Karray. Applying hybrid machine learning algorithms to assess customer risk-adjusted revenue in the financial industry. Electronic Commerce Research and Applications 56 (2022) 101202. doi:https://doi.org/10.1016/j.elerap.2022.101202.

[20] N. Tasgetiren, U. Tigrak, E. Bozan, G. Gul, E. Demirci, H. Saribiyik, M. S. Aktas. On the distributed software architecture of a data analysis workflow: A case study. Concurrency and Computation: Practice and Experience 34 (2022) e6522. doi:https://doi.org/10.1002/cpe.6522.

[21] K. K. Pandey, D. Shukla. Stratified remainder linear systematic sampling extension (srse) to improve computational efficiency of risk clustering algorithms. Reliability: Theory & Applications 16 (2021) 239–257. doi:10.24412/1932-2321-2021-465-239-257.

[22] I. Singh, N. Kumar, K. Srinivasa, S. Maini, U. Ahuja, S. Jain. A multi-level classification and modified pso clustering based ensemble approach for credit scoring. Applied Soft Computing 111 (2021) 107687. doi:https://doi.org/10.1016/j.asoc.2021.107687.

[23] D. Lazo, R. Calabrese, C. Bravo. The effects of customer segmentation, borrower behaviors and analytical methods on the performance of credit scoring models in the agribusiness sector. Journal of Credit Risk 16 (2020) 119–156. doi:10.21314/JCR.2020.272.

[24] A. Nazari, M. Mehregan, R. Tehrani. Credit scoring of bank depositor with clustering techniques for supply chain finance. International Journal of Supply Chain Management 8 (2019) 374–383.

[25] D. J. Philip, N. Sudarsanam, B. Ravindran. Improved insights on financial health through partially constrained hidden markov model clustering on loan repayment data. Data Base for Advances in Information Systems 49 (2018) 98–113. doi:10.1145/3242734.3242741.

[26] S. M. A. K. Firouzabadi, M. T. Taghavifard, S. K. Sajjadi, J. B. Soufi. A multi-objective optimisation model for assignment of service to bank customers by using data mining and simulation. International Journal of Electronic Customer Relation Management 11 (2018) 237–255. doi:https://doi.org/10.1504/IJECRM.2018.093766.

[27] E. T. Luthfi, F. W. Wibowo. Loan payment prediction using adaptive neuro fuzzy inference system. International Journal of Simulation: Systems, Science and Technology 18 (2017) 9.1–9.6. doi:10.5013/IJSSST.a.18.04.09.

[28] L. Wang, W. Zhang. A qualitatively analyzable two-stage ensemble model based on machine learning for credit risk early warning: Evidence from chinese manufacturing companies. Information Processing and Management 60 (2023). doi:https://doi.org/10.1016/j.ipm.2023.103267.

[29] P. Guerra, M. Castelli, N. Côrte-Real. Machine learning for liquidity risk modelling: A supervisory perspective. Economic Analysis and Policy 74 (2022) 175–187. doi:https://doi.org/10.1016/j.eap.2022.02.001.

[30] A. Petropoulos, V. Siakoulis, E. Stavroulakis. Towards an early warning system for sovereign defaults leveraging on machine learning methodologies. Intelligent Systems in Accounting, Finance and Management 29 (2022) 118–129. doi:https://doi.org/10.1002/isaf.1516.

[31] X. Wangsong. The default risk of bank customers based on embedded microprocessor wireless communication under the internet finance background. Mobile Information Systems 2022 (2022). doi:https://doi.org/10.1155/2022/8019033.

[32] H. Xie, Y. Shi. A big data technique for internet financial risk control. Mobile Information Systems 2022 (2022). doi:https://doi.org/10.1155/2022/9549868.

[33] X. Han. Construction of economic data management system based on bp neural network. Computational Intelligence and Neuroscience 2022 (2022). doi:https://doi.org/10.1155/2022/9036917.

[34] L.-L. Yin, Y.-W. Qin, Y. Hou, Z.-J. Ren. A convolutional neural network-based model for supply chain financial risk early warning. Computational Intelligence and Neuroscience 2022 (2022). doi:https://doi.org/10.1155/2022/7825597.

[35] W. Xie. Study on enterprise financial risk prevention and early warning system based on blockchain technology. Mobile Information Systems 2022 (2022). doi:https://doi.org/10.1155/2022/4435296.

[36] B. Huang, X. Yao, Y. Luo, J. Li. Improving financial distress prediction using textual sentiment of annual reports. Annals of Operations Research (2022) 1–28. doi:https://doi.org/10.1007/s10479-022-04633-3.

[37] L. Xu, W. Chen, S. Wang, B. S. Mohammed, R. Lakshmana Kumar. Analysis on risk awareness model and economic growth of finance industry. Annals of Operations Research (2022) 1–23. doi:https://doi.org/10.1007/s10479-021-04516-z.

[38] L. Tong, G. Tong. A novel financial risk early warning strategy based on decision tree algorithm. Scientific Programming 2022 (2022) 1–10. doi:https://doi.org/10.1155/2022/4648427.

[39] G. Yang. Research on financial credit evaluation and early warning system of internet of things driven by computer-aided technology. Computer-Aided Design and Applications 19 (2022) 158–169. doi:10.14733/cadaps.2022.S6.158-169.

[40] M. Jacobs. Validation of corporate probability of default models considering alternative use cases. International Journal of Financial Studies 9 (2021). doi:https://doi.org/10.3390/ijfs9040063.

[41] L. Zhu, M. Li, N. Metawa. Financial risk evaluation z-score model for intelligent iot-based enterprise. Information Processing and Management 58 (2021). doi:https://doi.org/10.1016/j.ipm.2021.102692.

[42] A. Aytaç Emin, B. Dalgıç, T. Azrak. Constructing a banking fragility index for islamic banks: definition impact on the predictive power of an early warning system. Applied Economics Letters 28 (2021) 1589–1593. doi:10.1080/13504851.2020.1834497.

[43] W. Zhang, R.-S. Chen, Y.-C. Chen, S.-Y. Lu, N. Xiong, C.-M. Chen. An effective digital system for intelligent financial environments. IEEE Access 7 (2019) 155965–155976. doi:10.1109/ACCESS.2019.2943907.

[44] M. Pompella, A. Dicanio. Ratings based inference and credit risk: Detecting likely-to-fail banks with the pc-mahalanobis method. Economic Modelling 67 (2017) 34–44. doi:https://doi.org/10.1016/j.econmod.2016.08.023.

[45] E. Berlinger. Implicit rating: A potential new method to alert crisis on the interbank lending market. Finance Research Letters 21 (2017) 277–283. doi:https://doi.org/10.1016/j.frl.2016.11.010.

[46] DOMO, What is hybrid machine learning?, 2023. URL: https://www.domo.com/glossary/what-is-hybrid-machine-learning, accessed: February 2023.

[47] J. vom Brocke, A. Hevner, A. Maedche. Introduction to design science research (2020) 1–13. doi:https://doi.org/10.1007/978-3-030-46781-4_1.

[48] K. Peffers, C. Tuunanen, Tuure Gengler, , M. Rossi, W. Hui, V. Virtanen, J. Bragge. The design science research process: A model for producing and presenting information systems research. Proceedings of First International Conference on Design Science Research in Information Systems and Technology DESRIST (2006) 83–106. doi:https://doi.org/10.48550/arXiv.2006.02763.

[49] K. Berwind, M. Bornschlegl, M. Hemmje, M. Kaufmann. Towards a cross industry standard process to support big data applications in virtual research environments (2017). URL: https://www.cerc-conf.eu/wp-content/uploads/2018/06/CERC-2016-proceedings.pdf.

[50] R. Bootcamps, What is data mining? a beginner's guide, 2022. URL: https://bootcamp.rutgers.edu/blog/what-is-data-mining/#what, accessed: September 2023.

[51] H. Salih, M. Basna, A. M. Abdulazeez. A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining 2 (2021) 20–30. doi:https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032.

[52] M. Greenacre, P. J. F. Groenen, T. Hastie, A. Iodice D'Enza, A. Markos, E. Tuzhilina. Principal component analysis. Nature Reviews Methods Primers 2 (2022). doi:https://doi.org/10.1038/s43586-022-00184-w.

[53] S. Kamperis, Principal component analysis limitations and how to overcome them, 2021. URL: https://ekamperi.github.io/mathematics/2021/02/23/pca-limitations.html, accessed: September 2023.

[54] Keboola, A guide to principal component analysis (pca) for machine learning, 2022. URL: https://www.keboola.com/blog/pca-machine-learning, accessed: September 2023.

[55] M. Syakur, B. Khusnul Khotimah, E. Rohman, B. Dwi Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP Conference Series: Materials Science and Engineering 336 (2018). doi:10.1088/1757-899X/336/1/012017.

[56] Towards Data Science, Clustering using k-means with implementation, 2020. URL: https://towardsdatascience.com/clustering-using-k-means-with-implementation-40988620a973, accessed: May 2023.

[57] C. Mengyao. Introduction to the k-means clustering algorithm based on the elbow method. Geoscience and Remote Sensing 3 (2020) 9–16. doi:http://dx.doi.org/10.23977/geors.2020.030102.

[58] Geeks for Geeks, Difference between k-means and dbscan clustering, 2022. URL: https://www.geeksforgeeks.org/difference-between-k-means-and-dbscan-clustering/, accessed: March 2023.

[59] M. Hahsler, M. Piekenbrock, D. Doran. Dbscan: Fast density-based clustering with r. Journal of Statistical Software 91 (2019) 1–30. doi:10.18637/jss.v091.i01.

[60] Medium, Visualizing clustering algorithms: K-means and dbscan, 2023. URL: https://levelup.gitconnected.com/visualizing-clustering-algorithms-k-means-and-dbscan-c4ce62de23c1, accessed: May 2023.

[61] P. I. Nakagawa, L. F. Pires, J. L. R. Moreira, L. O. Bonino da Silva Santos, F. Bukhsh. Semantic description of explainable machine learning workflows for improving trust. Appl. Sci. 11 (2021). doi:https://doi.org/10.3390/app112210804.

[62] S. Lundberg, S.-I. Lee. A unified approach to interpreting model predictions (2017) 4768—-4777. doi:10.5555/3295222.3295230.

[63] L. S. Shapley. A value for n-person games 2 (1953) 307–318. doi:https://doi.org/10.1515/9781400881970-018.

[64] Financial Times, 2023. URL: https://www.ft.com/, accessed: February 2023.

[65] Google News, 2023. URL: https://news.google.com/, accessed: February 2023.

[66] K. Wagstaff. Clustering with missing values: No imputation required (2004) 649–658. doi:https://doi.org/10.1007/978-3-642-17103-1_61.

[67] Towards Data Science, 2021. URL: https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10, accessed: April 2023.

[68] R. Nau, The logarithm transformation, 2019. URL: https://people.duke.edu/~rnau/411log.htm, accessed: April 2023.

[69] Machine Learning Mastery, Introduction to dimensionality reduction for machine learning, 2020. URL: https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/, accessed: May 2023.

[70] S. Kumar, I. Chong. Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. International Journal of Environmental Research and Public Health 15 (2018) 2907. doi:https://doi.org/10.3390/ijerph15122907.

[71] Analytics Vidhya, Feature engineering: Scaling, normalization and standardization, 2020. URL: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/, accessed: May 2023.

[72] scikit learn, sklearn.cluster.kmeans, 2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html, accessed: May 2023.

[73] Z. Akbari, R. Unland. Automated determination of the input parameter of dbscan based on outlier detection (2016) 280–291. doi:https://doi.org/10.1007/978-3-319-44944-9_24.

[74] Medium, How to measure clustering performances when there are no ground truth?, 2020. URL: https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c, accessed: June 2023.

[75] Medium, Performance metrics in machine learning — part 3: Clustering, 2021. URL: https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-d69550662dc6, accessed: June 2023.

A

APPENDIX A: SYSTEMATIC LITERATURE REVIEW

Table A.1: Summary of the articles examined in the SLR and their main features

Topic	Author	Settings	Main Purpose	Data-Driven Technique[1]	Metrics for evaluation[2]
CS	Kaminskyi et al. [17]	Online Lending	Customer Segmentation using a scoring approach and creation of an indicator of expected return of a borrower to determine a marketing solution	Whale curve for segmentation based on cumulative income, LR for creating scoring construction and SoftMax for generating vectors of probabilities of belonging to segments	–
CS	Jadwal, P.K., Jain, S., Pathak, S., Agarwal, B. [18]	Peer-to-Peer Lending	Customer Segmentation combined with SMOTE oversampling to reduce class imbalance and improve default prediction	Spectral clustering using the Elbow Method, benchmarking various SMOTE oversampling techniques on the clusters presenting a major of minor class instances, implementation of LR, SVM and KNN to test the effectiveness of the oversampling algorithms	G-Mean and F1 Score
CS	Machado, M.R., Karray, S. [19]	Peer-to-Peer Lending	Prediction of Customer's Risk-Adjusted Revenue (RAR) integrating Customer Segmentation	K-means using Elbow Method and DBSCAN for customer clustering, six regressors (AB, GB, DT, RF, SVM, and ANN) to predict RAR	EV, R^2, MAE, MSE, MedAE
CS	Tasgetiren, N., Tigrak, U., Bozan, E., Gul, G., Demirci, E., Saribiyik, H., Aktas, M.S. [20]	Banks	Hybrid distributed software architecture to segment customers and predict loan usage tendency	XGBoost and LightGBM for supervised learning, K-means, Bisecting K-means and Gaussian Mixture Model for unsupervised	Latency Test, Accuracy Test and Scalability Test for supervised Machine Learning workflows, Purity and Entropy metrics for unsupervised workflows

[1] Note: KNN refers to K-Nearest Neighbour, SVM is Support Vector Machine, CNN is Convolution Neural Networks, ANN is Artificial Neural Networks, AB is Adaboost, RF is Random Forest, GB is Gradient Boosting, DT is Decision Tree, NB is Naive Bayes, LR is Logistic Regression, SVDD is Support Vector Domain Description, ANFIS is Adaptive Neuro-Fizzy Inference Systems, CIF is Conditional Inference Trees, LSTM is Long Short-Term Memory

[2] Note: MSE is Mean Squared Error, RMSE is the Root Mean Squared Error, MAE is the Mean Absolute Error, ROC/AUC refers to the Area Under the ROC Curve, MedAE is Median Absolute Error

CS	Yuan, K., Chi, G., Zhou, Y., Yin, H. [16]	Banks and Lending Institutions	Default Prediction integrating customer clustering	K-Means Clustering for segmenting customers, SVDD for one-classification	AUC, G-Mean and Type-II Error
CS	Pandey, K.K., Shukla, D. [21]	Financial Instituions	Stratified Remainder Linear Systematic Sampling Extension (SRSE) to improve computational efficiency of risk clustering algorithms	Benchmarking the SSE-based clustering approach to the classical partitional K-means and K-means ++	Davies Bouldin score, Silhouette coefficient, Scattering Density between clusters Validity, Scattering Distance Validity and CPU Time
CS	Singh, In., Kumar, N., Srinivasa, K.G., Maini, S., Ahuja, U., Jain, S. [22]	Financial Instances	Classification of good and bad credit	Multi-Level Classification using 4 base classifiers (NN, KNN, SVM and RF), variation of Particle Swam Optimization algorithm for clustering the training dataset after the first classification to assign weights to the classifiers in different spacial regions	H measure, Precision, AUC, Recall, F1 Score and Accuracy
CS	Lazo, D., Calabrese, R., Bravo, C. [23]	Financing companies in the agrobusiness sector	Prediction of the probability of default of Chilean farmers	ClustOfVar algorithm for clustering features, LR, NN and for predicting the PD	AUC for the prediction, Mean Decrease Gini (MDG) and Mean Decrease Accuracy (MDA) to measure the importance of each variable
CS	Morandi, S., Mokharab Rafiei, F. [15]	Banks	Predicting banks credit risk level	Fuzzy C-Means (FCM) for clustering the clients, ANFIS to predict customers risk level, application of Fuzzy Interference System (FIS) on medium risk customers to identify "too risky" borrowers	Degree of sensitivity and degree of diagnosis
CS	Nazari, A., Mehregan, M., Tehrani, R. [24]	Banks and Credit Institutions	Customer Segmentation and credit scoring	Benchmarking K-Means, FCM and Sub-clustering techniques	LIFT and Silhouette Scores
CS	Philip, D.J., Sudarsanam, N., Ravindram, B. [25]	Financial Institutions	Clustering clients over time based on the repayment behaviour	Benchmarking normal and Partially Constrained Hidden Markov Models (PC-HMM) to cluster time series data	Measuring how the index of bank branch matches the clusters generated
CS	Firouzabadi, S.M.A.K., Taghavifard, M.T., Sajjadi, S.K., Soufi, J.B. [26]	Banks	Customer Segmentation for the optimal allocation of bank services	K-Means for clustering, suing Ward Method and Silhouette Score for optimal number of clusters	Distribution analysis of the variables within clusters
CS	Luthfi, E.T., Wibowo, F.W. [27]	Financial Institutions	Predicting loan payments using ANFIS	ANFIS	RMSE
EWS	Wang, L., Zhang, W. [28]	Chinese manufacturing companies	Predicting high credit risk companies	Two-stage ensemble model: the first stage used Grey Relational Analysis to select relevant indicators, the second stage implemented the Bagging Method to integrate 5 CNN models to make the prediction based on 5-fold cross-validation	Accuracy, Recall, Precision, F1 Score, G-Mean
EWS	Guerra, P., Castelli, M., Côrte-Real, N. [29]	Portoguese bank	Classification of banks' risk level	RF Classifier for feature selection, benchmarking of LR, SVM Classifier, NB Classifier, RF Classifier and XGBoost Classifier using train-test split, 10-fold cross validation and TPOT	F1 Score and Confusion Matrix

EWS	Petropoulos, A., Siakoulis, V., Stavroulakis, E. [30]	Sovreigns and governements	Creation of a sovreign rating system based on their risk of default	Benchmarking LR, SVM, NN, RF, CIF and XGBoost using train-test split, calibration to implement the credit rating system	AUC, Kolmogorov-Smirnov test, Youden Score, Negative Likelihood Ratio, Geometric Mean Balanced Accuracy for validating the prediction performance, SSE and Brier Score to validate the calibration
EWS	Wangsong, X. [31]	Internet Finance of Chinese banks	Use multimedia technology to design a bank customer default risk management system that measures and controls credit risk	–	Test performance of the system and users survey feedback
EWS	Xie, H., Shi, Y. [32]	Chinese companies with regards to Internet Finance	Creation of a Internet Financial Risk Control model using Big Data	Criteria Importance Through Intercriteria Correlation (CRITIC) method to determine the weight of the risks indexes and mathematical expression to calculate the probability of the financial risk	Error Rate, Accuracy, Control time
EWS	Han, X. [33]	Governments	Creation of Economic Data Management System	2 Back Propagation (2BP) Neural Network	–
EWS	Yin, L.L., Qin, Y., Hou, Y., Zhao, J.R. [34]	Supply Chain Finance	Creation of a Risk Early Warning Index System to predict the default risk	PCA for selection of Risk Early Warning Indicators, CNN for predicting the risk index value	Accuracy and Loss Function
EWS	Xie, W. [35]	Supply-chain Financing	Creation of a blockchain-based financial risk prevention system	Evolutionary game model to measure the endorsement fiduciary relationship between SMEs and major enterprises	–
EWS	Huang, B., Yao, X., Luo, Y., Li, J. [36]	Chinese firms	Predicting financial distress of companies using textual sentiment of annual reports	word2vec to create word vectors, NB, SVM, DT, KNN, CNN and LSTM for generating textual sentiment using train-test split, benchmarking five different classifiers (LR, NN, LS-SVM, RF and XGBoost) based on out-of-time and out-of-sample predictions using train/test and k-fold cross validation	AUC, Kolmogorov–Smirnov statistic , Brier score , precision, recall, F1 score, and total accuracy for out-of-time prediction
EWS	Xu, L., Chen, W., Wang, S., Mohammed, B.S., Lakshmana Kumar, R. [37]	Countries	Creation of a Machine Learning-based risk awareness model for financial crisis prediction based on a number of risk factors	Design, Solo and Easy Classification techniques	Accuracy Ratio
EWS	Tong, L., Tong, G. [38]	Enterprises	Predicting the Cash Flow risk	DT	Entropy, Split information and Gain ratio
EWS	Yang, G. [39]	Online Lending	Creation of Credit Risk Assessment model that predicts default risk	Benchmarking Random Forest, XGBoost and XGBoost Deep Forest	AUC
EWS	Jacobs, M. [40]	Publicly raded US companies	Predicting the Point-In-Time (PIT) and Through-The-Cycle (TTC) Probability of Default (PD) in one and three years horizons	LR	AUC, Hosmer-Lemeshow test, Akaike Information Criteriom (AIC), Singular Value Decomposition and Factor Contribution (for individual variables)

EWS	Zhu, L., Li, M., Metawa, N. [41]	IoT companies	Creation of a Risk Early Warning System composed of a Z-Score Model analysis to predict corporate finance and bankruptcy risk	Z-Score Model	–
EWS	Aytaç Emin, A., Dalgıç, B., Azrak, T. [42]	Islamic Banks	Creation of a new banking fragility index to improve the predictive power of an Early Warning System	Mathematical Formula for calculating the banking fragility index the	Predictive Power
EWS	Zhang, W., Chen, R.-S., Chen, Y.-C., Lu S-Y, Xiong, N., Chen, C.-M. [43]	Accouting departments of companies	Improving the company's financial model and evaluating the effectiveness of the internal control of the financial reporting system	–	Qualitative analysis
EWS	Pompella, M.,Dicanio, A. [44]	Publicly rated banks	Creation of an EWS that uses an accounting-based approach to identify high and low risk banks in order to test the validity of external ratings	Principal Commponent-Mahalanobis (PC-M) method	ROC curve
EWS	Berlinger, E. [45]	Interbank lending market	Introduction of a new indicator, called Implicit Rating (IR), for EWS	Mathematical formula for calculating the IR score	–
EWS in CS	Amato, A., Machado, M.R., Osterrieder, J., Rebelo Moreira, J.	Wholesale Banking	Customer clustering based on Early Warning signals	PCA and Correlation analysis for feature selection, K-Means and DBSCAN for clustering, SHAP Values for feature contribution analysis	Silhouette Score, Davies-Bouldin score, Calinski-Harabasz score, Users qualitative evaluation

B

APPENDIX B: DATA DICTIONARY

Table B.1: Dictionary of the data used

Data Field	Data Type	Description
outstanding_growth	float	Client's monthly growth of outstanding amount calculated considering logarithmic values
ead_growth	float	Client's monthly growth of EAD calculated considering logarithmic values
rwa_growth	float	Client's monthly growth of RWA calculated considering logarithmic values
expected_loss_growth	float	Client's monthly growth of expected loss calculated considering logarithmic values
allocated_limit_growth	float	Client's monthly growth of allocated limit calculated considering logarithmic values
tot_allocated_limit	float	Client's allocated limit recorded in the last month of activity
tot_outstanding	float	Client's outstanding amount recorded in the last month of activity
default_watchlist_status	boolean	Client's risk status recorded in the last month of activity
internal_triggers_counter	float	Client's average number of monthly internal triggers raised
external_triggers_counter	float	Client's average number of monthly external triggers raised
positive_articles	float	Client's average number of monthly positive articles raised
negative_articles	float	Client's average number of monthly negative triggers raised
rating_good_good_migration	boolean	Client's "good" to "good" internal rating status migration
rating_good_bad_migration	boolean	Client's "good" to "bad" internal rating status migration
rating_bad_good_migration	boolean	Client's "bad" to "good" internal rating status migration
rating_bad_bad_migration	boolean	Client's "bad" to "bad" internal rating status migration

ifrs_good_good_migration	boolean	Client's "good" to "good" worst IFRS status migration
ifrs_good_bad_migration	boolean	Client's "good" to "bad" worst IFRS status migration
ifrs_bad_good_migration	boolean	Client's "bad" to "good" worst IFRS status migration
ifrs_bad_bad_migration	booleab	Client's "bad" to "bad" worst IFRS status migration

C

APPENDIX C: CLUSTERS EXPLORATION

C.1. DESCRIPTIVE STATISTICS ANALYSIS

C.1.1. K-MEANS: UNCORRELATED DATASET

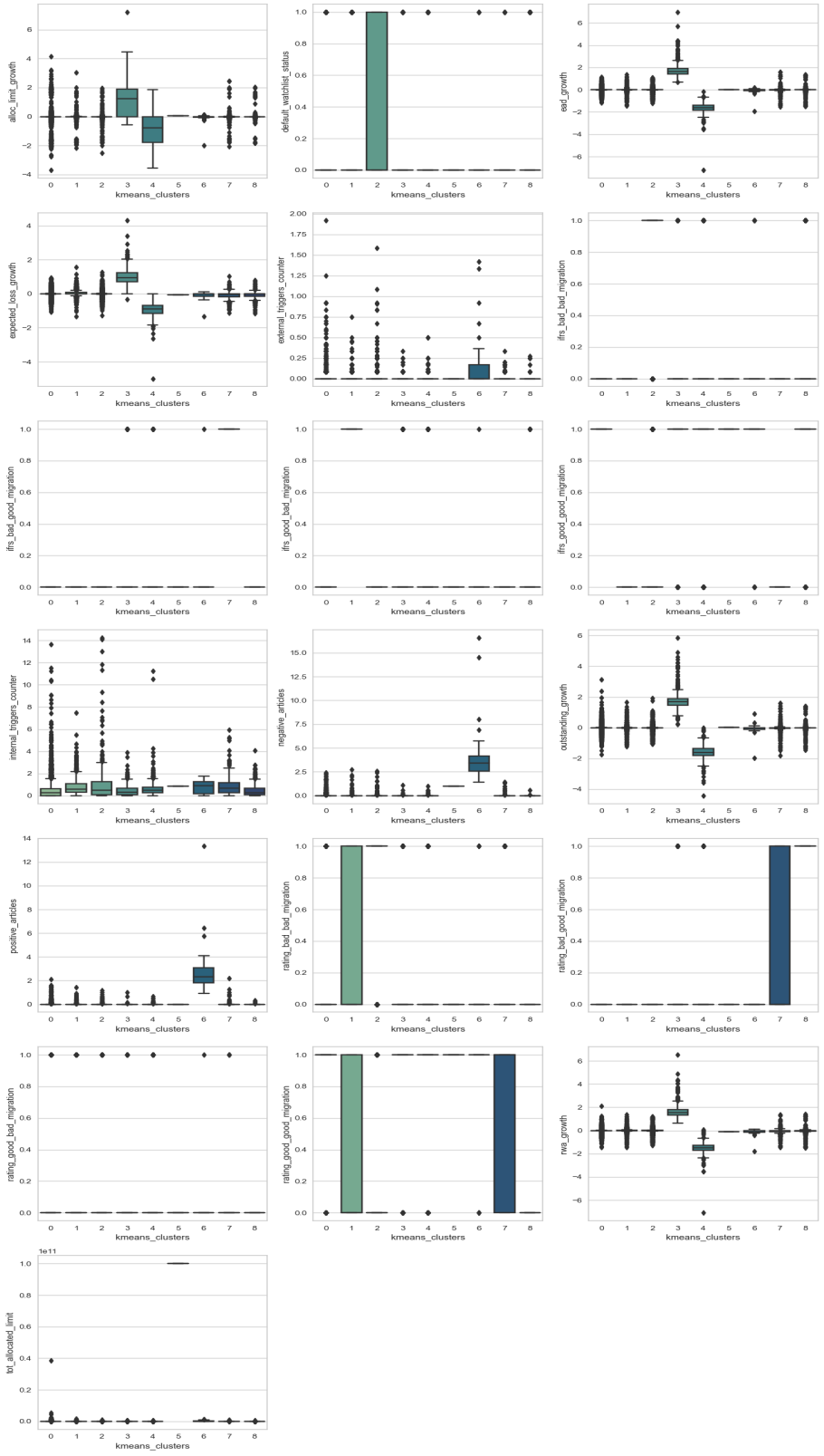


Figure C.1: Box plots of the features' values distribution for the uncorrelated dataset using K-Means

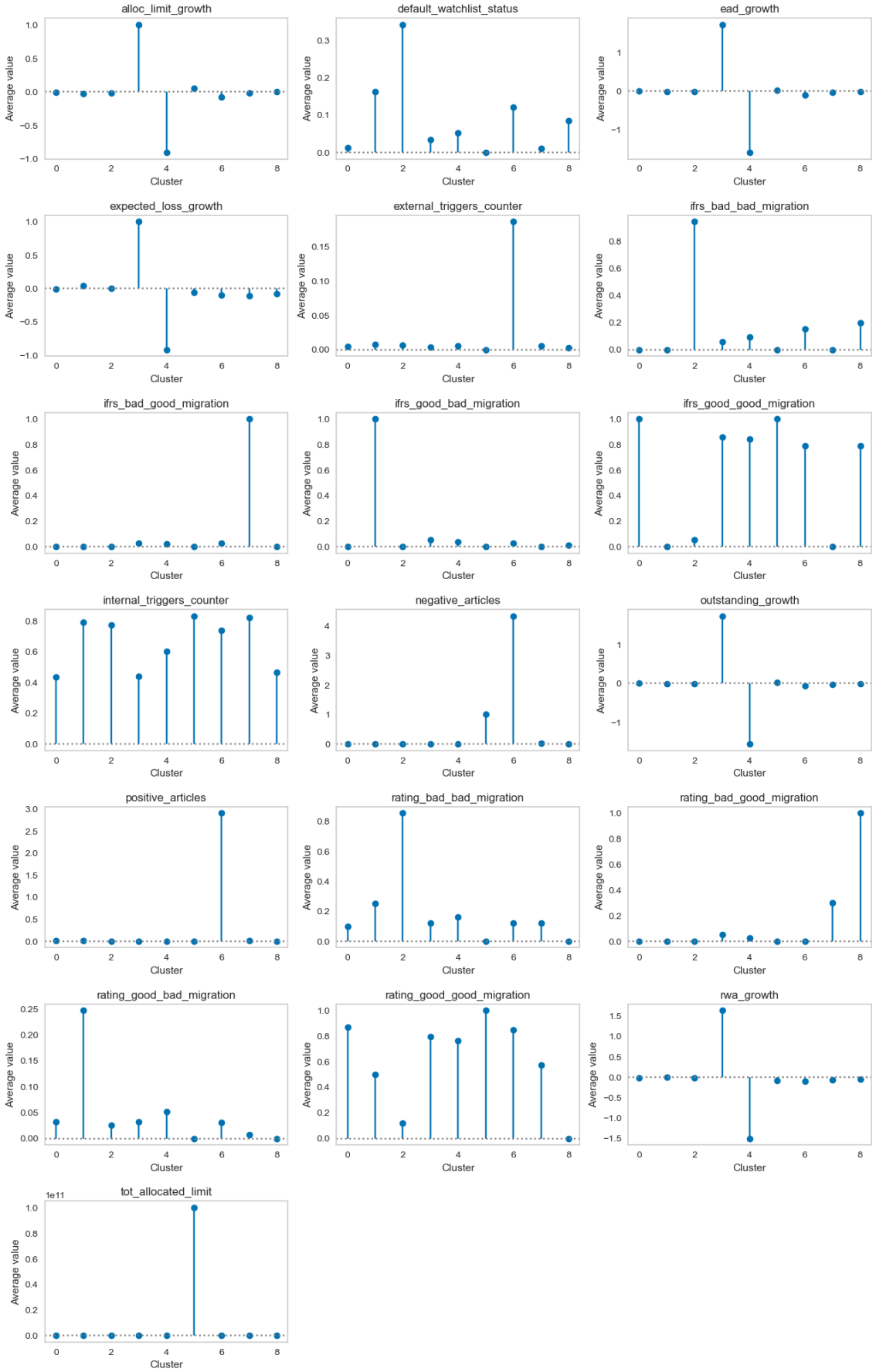


Figure C.2: Stem plots of the features' average values for the uncorrelated dataset using K-Means

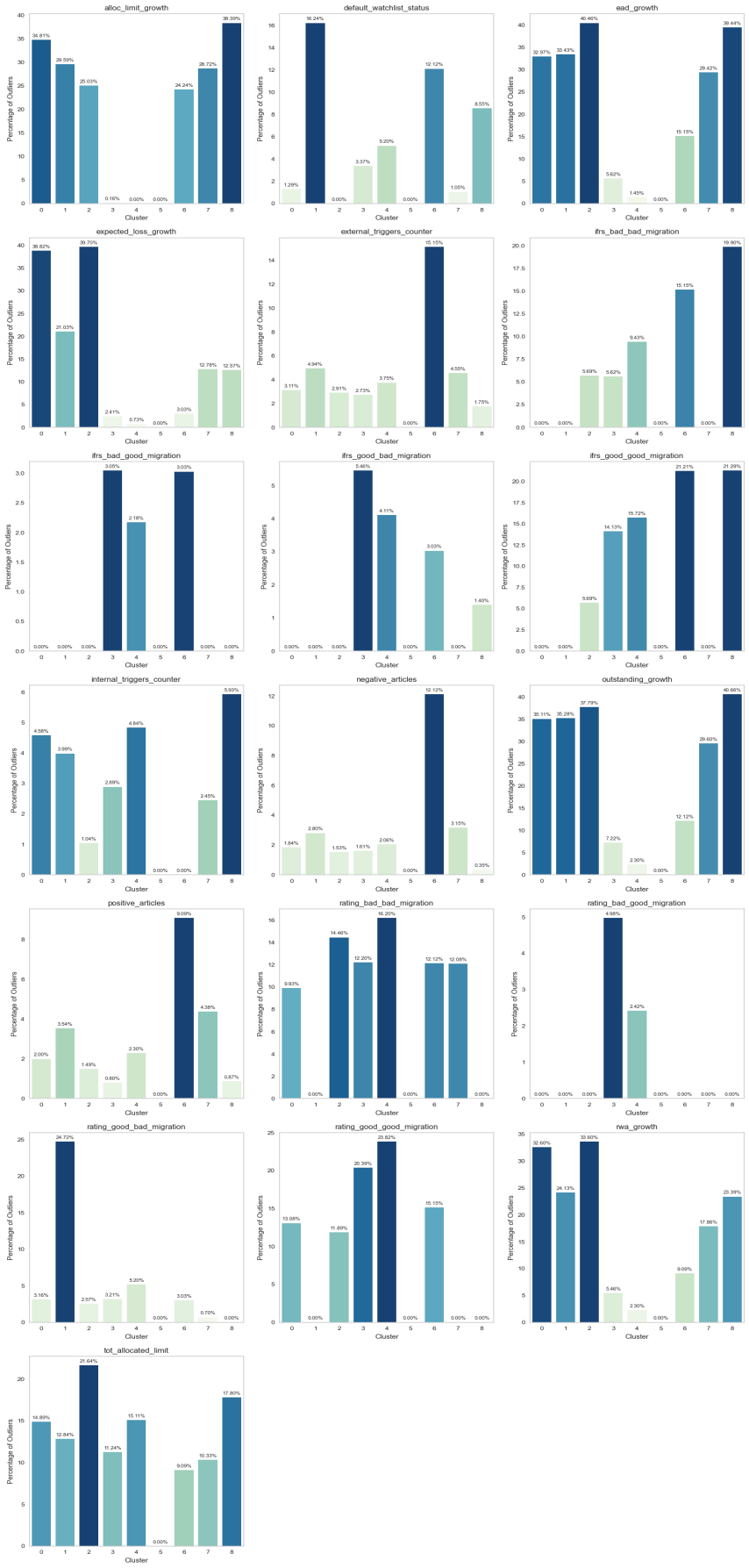


Figure C.3: Histogram of the features' outliers percentage for the uncorrelated dataset using K-Means

	kmeans_clusters	0	1	2	3	4	5	6	7	8
alloc_limit_growth	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	-5.113435e-04	-2.126709e-02	-1.349247e-02	1.009076e+00	-9.060943e-01	5.269615e-02	-7.506505e-02	-1.776035e-02	8.952034e-03
	std	2.449693e-01	2.769450e-01	2.082519e-01	1.014056e+00	9.050885e-01	NaN	3.532534e-01	3.446026e-01	2.908389e-01
	min	-3.684136e+00	-2.163446e+00	-2.503765e+00	-5.591863e-01	-3.566531e+00	5.269615e-02	-1.988580e+00	-2.062153e+00	-1.829584e+00
	25%	-2.659121e-03	-6.852745e-03	-1.716470e-03	0.000000e+00	-1.776213e+00	5.269615e-02	-3.320900e-02	-1.618742e-02	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	1.248252e+00	-7.827868e-01	5.269615e-02	0.000000e+00	-4.571550e-04	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	1.878854e+00	0.000000e+00	5.269615e-02	1.548440e-04	0.000000e+00	0.000000e+00
	max	4.144653e+00	3.043027e+00	1.976827e+00	7.213174e+00	1.867916e+00	5.269615e-02	1.418015e-01	2.445900e+00	2.030246e+00
default_watchlist_status	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	1.286860e-02	1.623616e-01	3.418863e-01	3.370787e-02	5.199516e-02	0.000000e+00	1.212121e-01	1.050788e-02	8.551483e-02
	std	1.127112e-01	3.689184e-01	4.744240e-01	1.806212e-01	2.221516e-01	NaN	3.314340e-01	1.020574e-01	2.798906e-01
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	max	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
ead_growth	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	-2.948399e-03	-9.338212e-03	-1.023264e-02	1.732784e+00	-1.597053e+00	2.501369e-02	-9.623506e-02	-2.865298e-02	-2.052924e-02
	std	1.053851e-01	1.816005e-01	1.872553e-01	5.776585e-01	4.075664e-01	NaN	3.424284e-01	2.638969e-01	2.867225e-01
	min	-1.164360e+00	-1.453764e+00	-1.207585e+00	6.819838e-01	-7.213263e+00	2.501369e-02	-1.914533e+00	-1.544058e+00	-1.467960e+00
	25%	-5.967929e-03	-1.094196e-02	-5.955094e-03	1.425174e+00	-1.842835e+00	2.501369e-02	-7.733136e-02	-2.705618e-02	-3.327508e-03
	50%	0.000000e+00	0.000000e+00	0.000000e+00	1.655858e+00	-1.595409e+00	2.501369e-02	0.000000e+00	-2.582768e-03	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	1.919445e+00	-1.369603e+00	2.501369e-02	0.000000e+00	0.000000e+00	0.000000e+00
	max	1.129929e+00	1.366874e+00	1.109643e+00	6.982210e+00	-1.900892e-01	2.501369e-02	1.704391e-01	1.578760e+00	1.360768e+00
expected_loss_growth	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	-7.002846e-03	3.898015e-02	-7.776350e-04	1.000042e+00	-9.148574e-01	-6.080926e-02	-9.485070e-02	-1.048801e-01	-8.209199e-02
	std	1.281545e-01	1.940097e-01	1.833031e-01	4.414597e-01	3.779181e-01	NaN	2.566452e-01	2.278505e-01	2.101051e-01
	min	-1.081595e+00	-1.343310e+00	-1.285974e+00	-3.378251e-01	-4.992291e+00	-6.080926e-02	-1.339992e+00	-1.125652e+00	-1.153788e+00
	25%	-1.003577e-02	-3.390425e-03	-5.644573e-03	7.184695e-01	-1.155083e+00	-6.080926e-02	-1.442523e-01	-1.829485e-01	-1.523452e-01
	50%	0.000000e+00	0.000000e+00	0.000000e+00	9.438511e-01	-8.821952e-01	-6.080926e-02	0.000000e+00	-3.646431e-02	0.000000e+00
	75%	0.000000e+00	7.888592e-02	9.461804e-03	1.250678e+00	-6.893926e-01	-6.080926e-02	6.421014e-03	0.000000e+00	0.000000e+00
	max	9.484018e-01	1.555300e+00	1.259187e+00	4.307961e+00	-1.238588e-02	-6.080926e-02	1.196998e-01	1.036157e+00	7.889233e-01
external_triggers_counter	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	5.196550e-03	7.648440e-03	6.894810e-03	3.742886e-03	5.505478e-03	0.000000e+00	1.866085e-01	5.692527e-03	2.802898e-03
	std	4.183238e-02	4.344551e-02	5.940450e-02	2.580505e-02	3.481798e-02	NaN	3.743575e-01	2.896535e-02	2.364801e-02
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.666667e-01	0.000000e+00	0.000000e+00
	max	1.916667e+00	7.500000e-01	1.583333e+00	3.333333e-01	5.000000e-01	0.000000e+00	1.416667e+00	3.333333e-01	2.727273e-01
ifrs_bad_bad_migration	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	0.000000e+00	0.000000e+00	9.431345e-01	5.617978e-02	9.431681e-02	0.000000e+00	1.515152e-01	0.000000e+00	1.989529e-01
	std	0.000000e+00	0.000000e+00	2.316255e-01	2.304536e-01	2.924458e-01	NaN	3.641095e-01	0.000000e+00	3.995613e-01
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	max	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00
ifrs_bad_good_migration	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	0.000000e+00	0.000000e+00	0.000000e+00	3.049759e-02	2.176542e-02	0.000000e+00	3.030303e-02	1.000000e+00	0.000000e+00
	std	0.000000e+00	0.000000e+00	0.000000e+00	1.720902e-01	1.460050e-01	NaN	1.740777e-01	0.000000e+00	0.000000e+00
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00
	max	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00
ifrs_good_bad_migration	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	0.000000e+00	1.000000e+00	0.000000e+00	5.457464e-02	4.111245e-02	0.000000e+00	3.030303e-02	0.000000e+00	1.396161e-02
	std	0.000000e+00	0.000000e+00	0.000000e+00	2.273306e-01	1.986704e-01	NaN	1.740777e-01	0.000000e+00	1.174340e-01
	min	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	max	0.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00
ifrs_good_good_migration	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	1.000000e+00	0.000000e+00	5.686546e-02	8.587480e-01	8.428053e-01	1.000000e+00	7.878788e-01	0.000000e+00	7.870855e-01
	std	0.000000e+00	0.000000e+00	2.316255e-01	3.485612e-01	3.642045e-01	NaN	4.151488e-01	0.000000e+00	4.097254e-01
	min	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00
	50%	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00
	75%	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00
	max	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00

Figure C.4: Table of features statistics for the uncorrelated dataset using K-Means

internal_triggers_counter	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	4.380550e-01	7.944337e-01	7.740309e-01	4.426287e-01	6.041351e-01	8.333333e-01	7.382155e-01	8.217141e-01	4.667829e-01
	std	6.206979e-01	6.870843e-01	9.634490e-01	4.679847e-01	7.185136e-01	NaN	5.910911e-01	7.816268e-01	5.477470e-01
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	8.333333e-01	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	3.333333e-01	8.333333e-02	8.333333e-02	2.500000e-01	8.333333e-01	1.666667e-01	2.500000e-01	8.333333e-02
	50%	2.500000e-01	5.833333e-01	5.000000e-01	3.333333e-01	5.000000e-01	8.333333e-01	9.090909e-01	6.666667e-01	2.727273e-01
	75%	6.250000e-01	1.083333e+00	1.250000e+00	6.666667e-01	7.777778e-01	8.333333e-01	1.250000e+00	1.166667e+00	6.666667e-01
	max	1.366667e+01	7.500000e+00	1.425000e+01	3.916667e+00	1.125000e+01	8.333333e-01	1.750000e+00	5.916667e+00	4.083333e+00
negative_articles	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	6.933912e-03	1.502404e-02	7.587841e-03	4.414125e-03	6.146715e-03	1.000000e+00	4.336838e+00	1.774664e-02	1.192554e-03
	std	7.886743e-02	1.378929e-01	9.866842e-02	5.232746e-02	5.505309e-02	NaN	3.236920e+00	1.291610e-01	2.471738e-02
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.416667e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	2.583333e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	3.400000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	4.166667e+00	0.000000e+00	0.000000e+00
	max	2.416667e+00	2.750000e+00	2.583333e+00	1.083333e+00	1.000000e+00	1.000000e+00	1.658333e+01	1.416667e+00	5.833333e-01
outstanding_growth	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	5.556044e-04	-8.287505e-03	-9.194662e-03	1.734263e+00	-1.562323e+00	2.132921e-02	-6.838825e-02	-2.491046e-02	-1.264281e-02
	std	1.332159e-01	2.024534e-01	2.016956e-01	5.638277e-01	4.080489e-01	NaN	3.935813e-01	2.957271e-01	2.741874e-01
	min	-1.758777e+00	-1.226279e+00	-1.052688e+00	2.302585e-01	-4.424374e+00	2.132921e-02	-1.977546e+00	-1.794374e+00	-1.444082e+00
	25%	-6.212591e-03	-1.555962e-02	-1.018962e-02	1.448369e+00	-1.821658e+00	2.132921e-02	-1.115616e-01	-2.922822e-02	-2.280118e-03
	50%	0.000000e+00	0.000000e+00	0.000000e+00	1.691071e+00	-1.603966e+00	2.132921e-02	-1.221757e-02	-1.428183e-03	0.000000e+00
	75%	0.000000e+00	8.892790e-05	1.215136e-03	1.902922e+00	-1.344019e+00	2.132921e-02	0.000000e+00	5.892208e-03	0.000000e+00
	max	3.142006e+00	1.667771e+00	1.928017e+00	5.858981e+00	-1.976362e-03	2.132921e-02	9.008504e-01	1.578760e+00	1.412035e+00
positive_articles	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	6.857026e-03	1.032912e-02	4.564255e-03	3.222433e-03	5.159210e-03	0.000000e+00	2.908402e+00	1.743997e-02	1.541594e-03
	std	6.969667e-02	7.571679e-02	4.962569e-02	4.879516e-02	4.144109e-02	NaN	2.227307e+00	1.296205e-01	1.835951e-02
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	9.090909e-01	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.818182e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.333333e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.083333e+00	0.000000e+00	0.000000e+00
	max	2.083333e+00	1.416667e+00	1.166667e+00	1.000000e+00	6.666667e-01	0.000000e+00	1.333333e+01	2.166667e+00	3.000000e-01
rating_bad_bad_migration	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	9.926280e-02	2.538745e-01	8.554092e-01	1.219904e-01	1.620314e-01	0.000000e+00	1.212121e-01	1.208406e-01	0.000000e+00
	std	2.990242e-01	4.353874e-01	3.517488e-01	3.275377e-01	3.687026e-01	NaN	3.314340e-01	3.262278e-01	0.000000e+00
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	max	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00
rating_bad_good_migration	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	0.000000e+00	0.000000e+00	0.000000e+00	4.975923e-02	2.418380e-02	0.000000e+00	0.000000e+00	2.977233e-01	1.000000e+00
	std	0.000000e+00	0.000000e+00	0.000000e+00	2.176218e-01	1.537124e-01	NaN	0.000000e+00	4.576581e-01	0.000000e+00
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
	max	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
rating_good_bad_migration	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	3.155717e-02	2.472325e-01	2.565881e-02	3.210273e-02	5.199516e-02	0.000000e+00	3.030303e-02	7.005254e-03	0.000000e+00
	std	1.748236e-01	4.315623e-01	1.581427e-01	1.764146e-01	2.221516e-01	NaN	1.740777e-01	8.347685e-02	0.000000e+00
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	max	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00
rating_good_good_migration	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	8.691800e-01	4.988930e-01	1.189320e-01	7.961477e-01	7.617896e-01	1.000000e+00	8.484848e-01	5.744308e-01	0.000000e+00
	std	3.372143e-01	5.001834e-01	3.237647e-01	4.031842e-01	4.262463e-01	NaN	3.641095e-01	4.948625e-01	0.000000e+00
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00
	50%	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00
	75%	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00
	max	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00
rwa_growth	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	-7.930330e-03	1.946369e-03	-1.992409e-02	1.636026e+00	-1.497161e+00	-8.889946e-02	-1.014139e-01	-6.320871e-02	-4.652017e-02
	std	1.242902e-01	1.952515e-01	2.135015e-01	5.551693e-01	4.113844e-01	NaN	3.276536e-01	2.569591e-01	2.982093e-01
	min	-1.461414e+00	-1.441692e+00	-1.280670e+00	6.466251e-01	-7.091768e+00	-8.889946e-02	-1.811440e+00	-1.470188e+00	-1.480095e+00
	25%	-1.385437e-02	-1.237714e-02	-1.852171e-02	1.324622e+00	-1.715876e+00	-8.889946e-02	-1.289491e-01	-9.455136e-02	-5.098050e-02
	50%	0.000000e+00	0.000000e+00	0.000000e+00	1.553476e+00	-1.494294e+00	-8.889946e-02	-8.315608e-03	-1.752995e-02	0.000000e+00
	75%	0.000000e+00	2.709696e-02	3.366265e-03	1.809197e+00	-1.284966e+00	-8.889946e-02	0.000000e+00	2.266821e-04	0.000000e+00
	max	2.123463e+00	1.377630e+00	1.188894e+00	6.520162e+00	6.236095e-02	-8.889946e-02	1.309434e-01	1.367713e+00	1.397845e+00
tot_allocated_limit	count	1.546400e+04	1.355000e+03	2.884000e+03	6.230000e+02	8.270000e+02	1.000000e+00	3.300000e+01	5.710000e+02	5.730000e+02
	mean	3.232632e+07	3.407285e+07	9.870618e+06	2.728841e+07	7.551897e+06	1.002000e+11	2.603657e+08	3.971939e+07	7.123964e+06
	std	3.311979e+08	1.070024e+08	4.478005e+07	6.155142e+07	4.041659e+07	NaN	3.825308e+08	9.274934e+07	3.584090e+07
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.002000e+11	0.000000e+00	0.000000e+00	0.000000e+00
	25%	3.000000e+04	0.000000e+00	0.000000e+00	5.456499e+05	0.000000e+00	1.002000e+11	1.638370e+05	1.278656e+05	1.032955e+04
	50%	7.836038e+05	1.515979e+06	1.308380e+03	4.839327e+06	1.000000e+00	1.002000e+11	4.000000e+07	5.539444e+06	1.600000e+05
	75%	1.880385e+07	2.501825e+07	5.487007e+05	2.634645e+07	2.000000e+06	1.002000e+11	4.347640e+08	3.632169e+07	1.150390e+06
	max	3.865231e+10	1.700000e+09	8.916667e+08	6.316271e+08	7.000012e+08	1.002000e+11	1.244000e+09	8.710000e+08	5.259501e+08

Figure C.5: Continuation of table of features statistics for the uncorrelated dataset using K-Means

C.1.2. K-MEANS: PCA DATASET

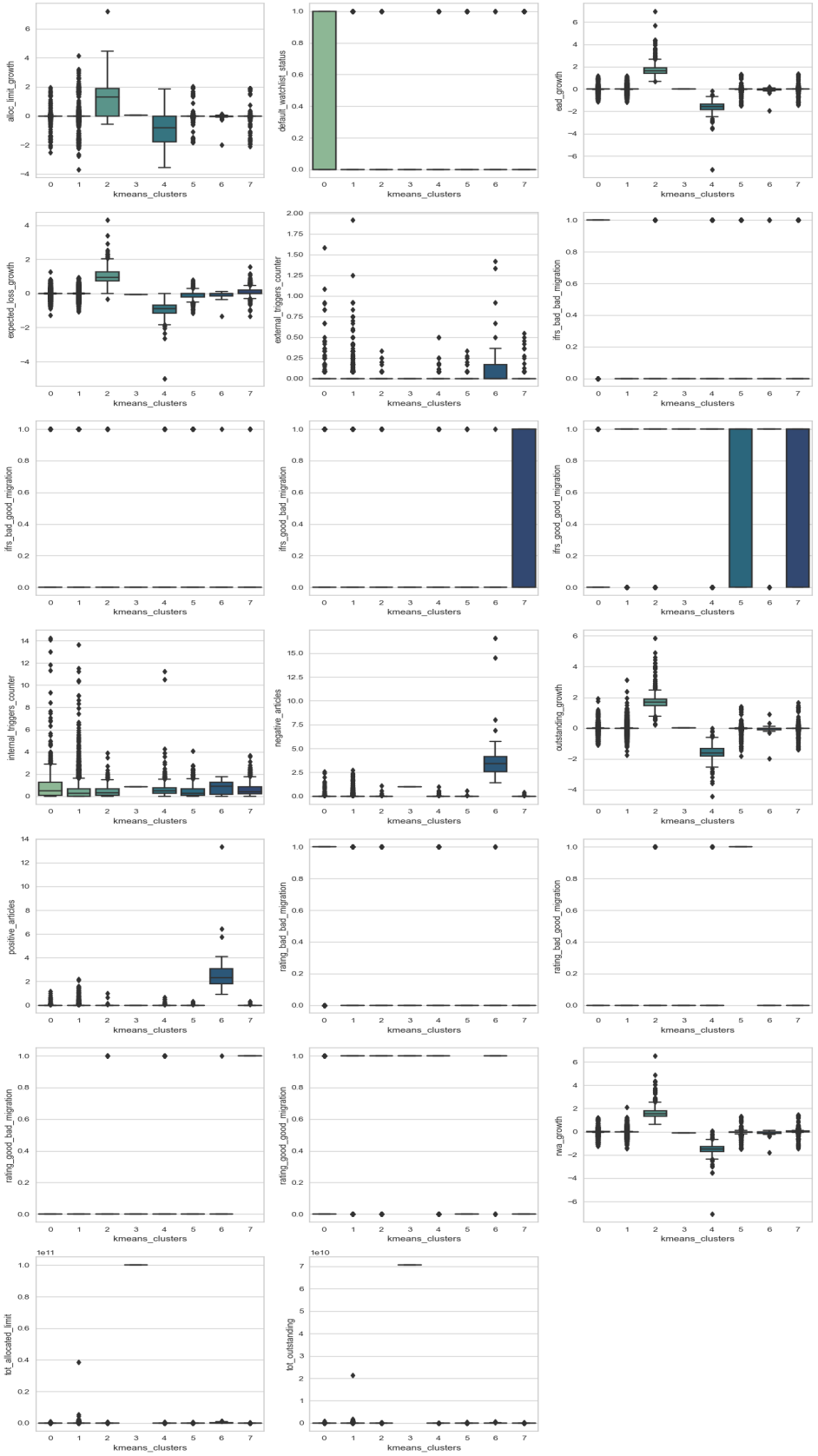


Figure C.6: Box plots of the features' values distribution for the PCA dataset using K-Means

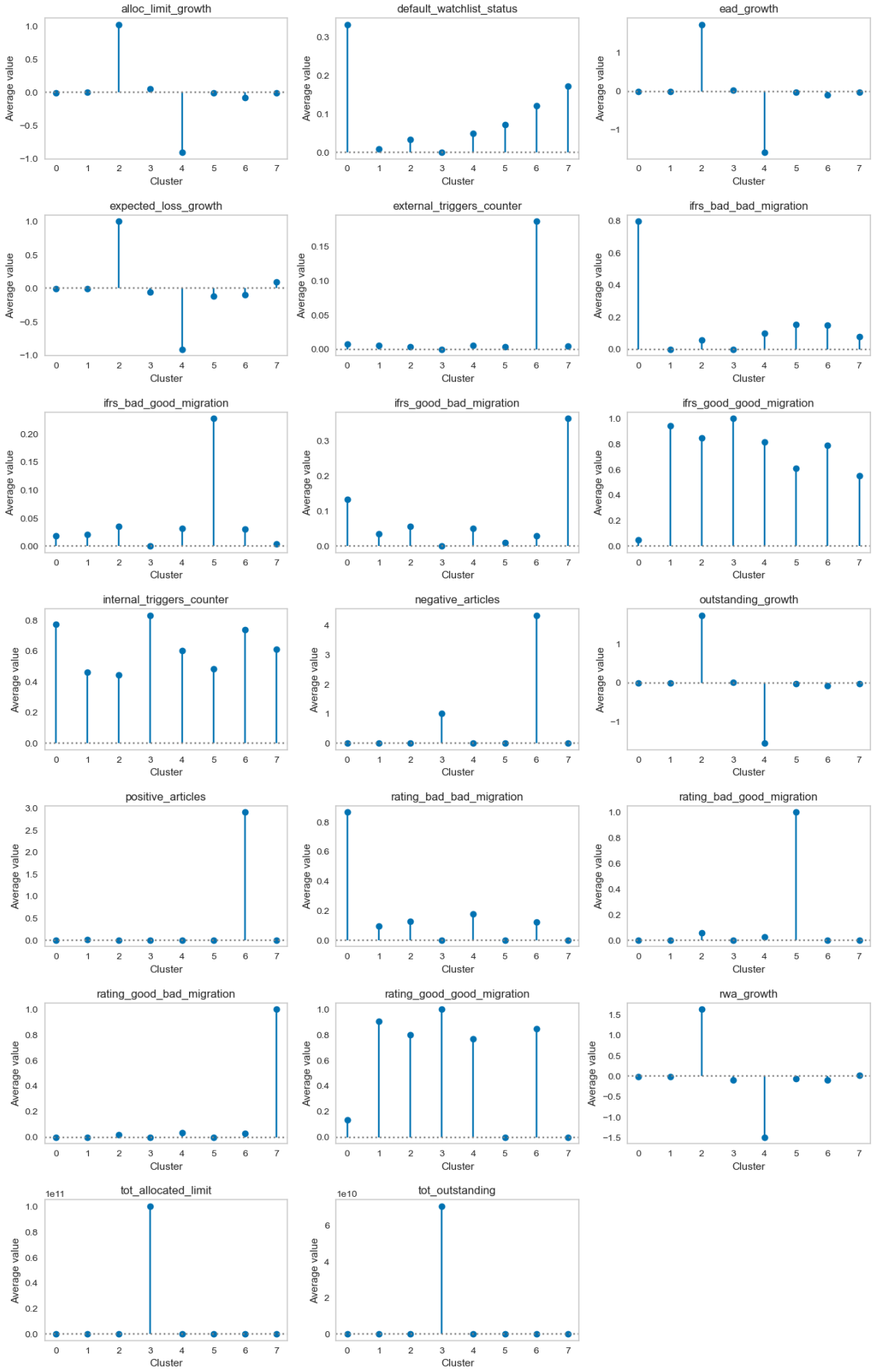


Figure C.7: Stem plots of the features' average values for the PCA dataset using K-Means

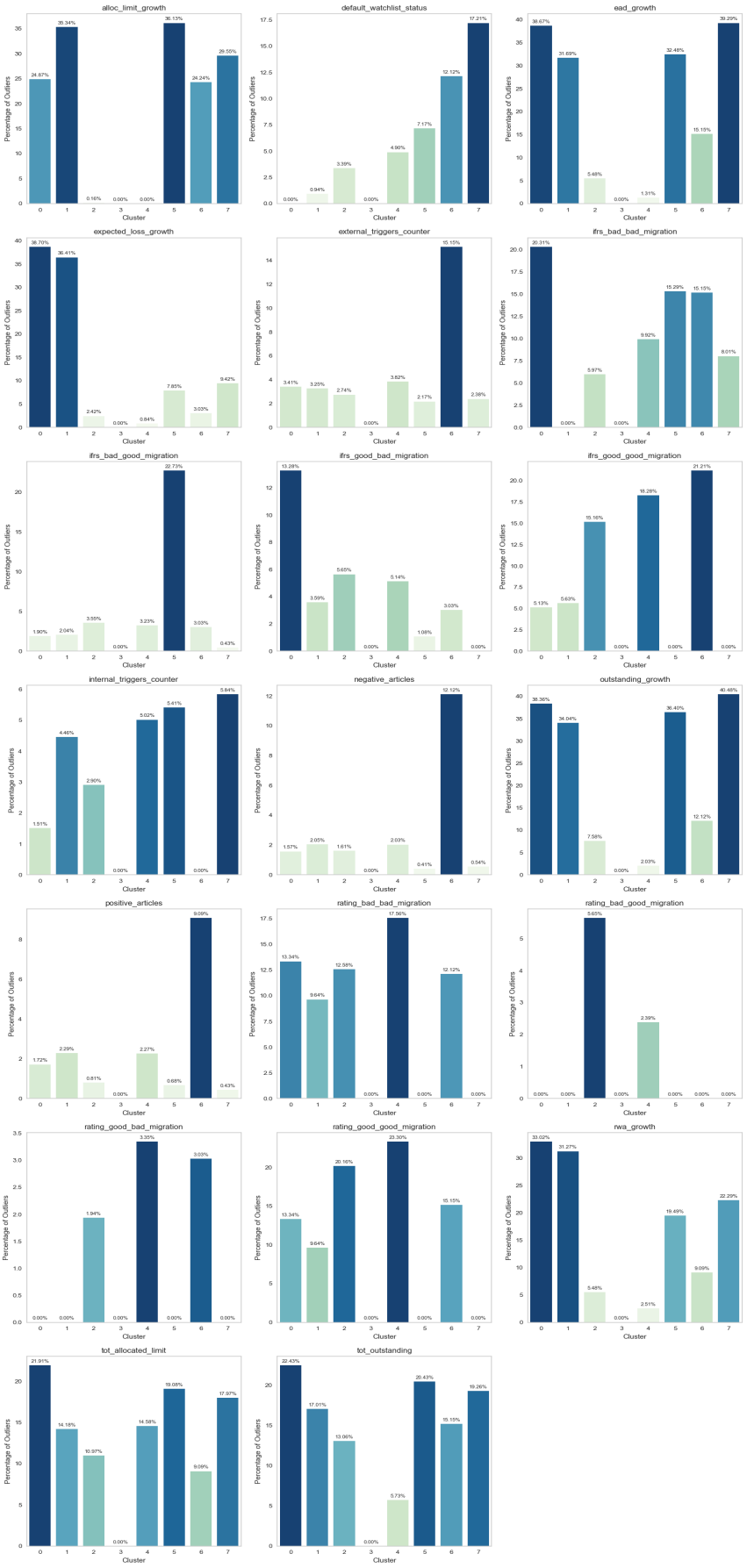


Figure C.8: Histogram of the features' outliers percentage for the PCA dataset using K-Means

	kmeans_clusters	0	1	2	3	4	5	6	7
alloc_limit_growth	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	-1.055388e-02	-1.807816e-03	1.029989e+00	5.269615e-02	-9.107297e-01	-6.556051e-03	-7.506505e-02	-9.910535e-03
	std	2.064787e-01	2.382438e-01	1.012876e+00	NaN	9.047080e-01	3.304457e-01	3.532534e-01	3.449766e-01
	min	-2.503765e+00	-3.684136e+00	-5.591863e-01	5.269615e-02	-3.566531e+00	-1.829584e+00	-1.988580e+00	-2.094994e+00
	25%	-2.582264e-03	-2.946038e-03	0.000000e+00	5.269615e-02	-1.777415e+00	-2.426300e-03	-3.320900e-02	-2.221392e-03
	50%	0.000000e+00	0.000000e+00	1.318969e+00	5.269615e-02	-7.992984e-01	0.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	1.882066e+00	5.269615e-02	0.000000e+00	0.000000e+00	1.548440e-04	0.000000e+00
	max	1.976827e+00	4.144653e+00	7.213174e+00	5.269615e-02	1.867916e+00	2.030246e+00	1.418015e-01	1.904219e+00
default_watchlist_status	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	3.323272e-01	9.392335e-03	3.387097e-02	0.000000e+00	4.898447e-02	7.171854e-02	1.212121e-01	1.720779e-01
	std	4.711187e-01	9.646090e-02	1.810431e-01	NaN	2.159646e-01	2.581960e-01	3.314340e-01	3.776526e-01
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	max	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
ead_growth	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	-7.460456e-03	-2.983048e-03	1.736573e+00	2.501369e-02	-1.588691e+00	-2.259423e-02	-9.623506e-02	-1.658619e-02
	std	1.768925e-01	1.070481e-01	5.774230e-01	NaN	4.137697e-01	2.744597e-01	3.424284e-01	2.471831e-01
	min	-1.100522e+00	-1.164360e+00	6.819838e-01	2.501369e-02	-7.213263e+00	-1.467960e+00	-1.914533e+00	-1.453764e+00
	25%	-6.188278e-03	-7.506397e-03	1.432093e+00	2.501369e-02	-1.838898e+00	-1.320724e-02	-7.733136e-02	-2.582768e-03
	50%	0.000000e+00	0.000000e+00	1.657859e+00	2.501369e-02	-1.590790e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	1.925586e+00	2.501369e-02	-1.355828e+00	0.000000e+00	0.000000e+00	0.000000e+00
	max	1.190000e+00	1.192942e+00	6.982210e+00	2.501369e-02	-1.900892e-01	1.340816e+00	1.704391e-01	1.366874e+00
expected_loss_growth	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	-2.692842e-03	-9.871776e-03	1.003783e+00	-6.080926e-02	-9.142864e-01	-1.115309e-01	-9.485070e-02	1.002794e-01
	std	1.685857e-01	1.281212e-01	4.397082e-01	NaN	3.758357e-01	2.218894e-01	2.566452e-01	2.349791e-01
	min	-1.285974e+00	-1.081595e+00	-3.378251e-01	-6.080926e-02	-4.992291e+00	-1.153788e+00	-1.339992e+00	-1.343310e+00
	25%	-5.762086e-03	-1.395837e-02	7.243502e-01	-6.080926e-02	-1.151856e+00	-2.091442e-01	-1.442523e-01	0.000000e+00
	50%	0.000000e+00	0.000000e+00	9.468600e-01	-6.080926e-02	-8.794221e-01	0.000000e+00	0.000000e+00	0.000000e+00
	75%	1.071599e-02	0.000000e+00	1.254651e+00	-6.080926e-02	-6.910118e-01	0.000000e+00	6.421014e-03	1.926078e-01
	max	1.259187e+00	9.484018e-01	4.307961e+00	-6.080926e-02	-1.238588e-02	7.889233e-01	1.196998e-01	1.555300e+00
external_triggers_counter	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	7.345043e-03	5.276991e-03	3.760997e-03	0.000000e+00	5.539264e-03	3.469061e-03	1.866085e-01	5.161190e-03
	std	5.833650e-02	4.143223e-02	2.586619e-02	NaN	3.471831e-02	2.645569e-02	3.743575e-01	4.116947e-02
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.666667e-01	0.000000e+00
	max	1.583333e+00	1.916667e+00	3.333333e-01	0.000000e+00	5.000000e-01	3.333333e-01	1.416667e+00	5.454545e-01
ifrs_bad_bad_migration	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	7.968609e-01	0.000000e+00	5.967742e-02	0.000000e+00	9.916368e-02	1.529093e-01	1.515152e-01	8.008658e-02
	std	4.023960e-01	0.000000e+00	2.370795e-01	NaN	2.990604e-01	3.601438e-01	3.641095e-01	2.715742e-01
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	max	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
ifrs_bad_good_migration	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	1.901600e-02	2.042360e-02	3.548387e-02	0.000000e+00	3.225806e-02	2.273342e-01	0.030303e-02	4.329004e-03
	std	1.366017e-01	1.414487e-01	1.851487e-01	NaN	1.767903e-01	4.193941e-01	1.740777e-01	6.568816e-02
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	max	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
ifrs_good_bad_migration	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	1.328101e-01	3.586737e-02	5.645161e-02	0.000000e+00	5.137395e-02	1.082544e-02	0.030303e-02	3.636364e-01
	std	3.394207e-01	1.859653e-01	2.309781e-01	NaN	2.208913e-01	1.035508e-01	1.740777e-01	4.813062e-01
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
	max	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
ifrs_good_good_migration	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	5.131301e-02	9.437090e-01	8.483871e-01	1.000000e+00	8.172043e-01	6.089310e-01	7.878788e-01	5.519481e-01
	std	2.206687e-01	2.304900e-01	3.589349e-01	NaN	3.867300e-01	4.883203e-01	4.151488e-01	4.975634e-01
	min	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00
	50%	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
	75%	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
	max	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00

Figure C.9: Table of features statistics for the PCA dataset using K-Means

internal_triggers_counter	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	7.762982e-01	4.612860e-01	4.471200e-01	8.333333e-01	6.025937e-01	4.836291e-01	7.382155e-01	6.128871e-01
	std	9.507068e-01	6.370679e-01	4.699537e-01	NaN	7.159002e-01	5.508194e-01	5.910911e-01	5.695183e-01
	min	0.000000e+00	0.000000e+00	0.000000e+00	8.333333e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	9.090909e-02	0.000000e+00	8.333333e-02	8.333333e-01	2.500000e-01	8.333333e-02	1.666667e-01	2.000000e-01
	50%	5.000000e-01	2.500000e-01	3.333333e-01	8.333333e-01	5.000000e-01	2.857143e-01	9.090909e-01	4.166667e-01
	75%	1.250000e+00	6.666667e-01	6.666667e-01	8.333333e-01	7.777778e-01	6.833333e-01	1.250000e+00	8.333333e-01
	max	1.425000e+01	1.366667e+01	3.916667e+00	8.333333e-01	1.125000e+01	4.083333e+00	1.750000e+00	3.666667e+00
negative_articles	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	7.977298e-03	8.331535e-03	4.435484e-03	1.000000e+00	6.073278e-03	1.037438e-03	4.336838e+00	9.920635e-04
	std	9.882742e-02	8.946753e-02	5.245323e-02	NaN	5.472692e-02	2.197642e-02	3.236920e+00	1.665516e-02
	min	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	1.416667e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	2.583333e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	3.400000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	4.166667e+00	0.000000e+00
	max	2.583333e+00	2.750000e+00	1.083333e+00	1.000000e+00	1.000000e+00	5.833333e-01	1.658333e+01	4.166667e-01
outstanding_growth	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	-5.666426e-03	6.366186e-04	1.735784e+00	2.132921e-02	-1.552086e+00	-2.276948e-02	-6.838825e-02	-1.494406e-02
	std	1.938164e-01	1.371235e-01	5.665127e-01	NaN	4.148266e-01	2.775835e-01	3.935813e-01	2.578295e-01
	min	-1.096528e+00	-1.758777e+00	2.302585e-01	2.132921e-02	-4.424374e+00	-1.794374e+00	-1.977546e+00	-1.400631e+00
	25%	-8.928188e-03	-8.062335e-03	1.450333e+00	2.132921e-02	-1.820446e+00	-1.123014e-02	-1.115616e-01	-2.221392e-03
	50%	0.000000e+00	0.000000e+00	1.691842e+00	2.132921e-02	-1.599625e+00	0.000000e+00	-1.221757e-02	0.000000e+00
	75%	8.723857e-04	0.000000e+00	1.903376e+00	2.132921e-02	-1.333371e+00	0.000000e+00	0.000000e+00	0.000000e+00
	max	1.928017e+00	3.142006e+00	5.858981e+00	2.132921e-02	-1.976362e-03	1.412035e+00	9.008504e-01	1.667771e+00
positive_articles	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	4.678289e-03	7.992897e-03	3.238025e-03	0.000000e+00	5.097571e-03	1.195309e-03	2.908402e+00	9.264725e-04
	std	4.855295e-02	7.596668e-02	4.891274e-02	NaN	4.119631e-02	1.617615e-02	2.227307e+00	1.541834e-02
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	9.090909e-01	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.818182e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.333333e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.083333e+00	0.000000e+00
	max	1.166667e+00	2.166667e+00	1.000000e+00	0.000000e+00	6.666667e-01	3.000000e-01	1.333333e+01	3.333333e-01
rating_bad_bad_migration	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	8.665862e-01	9.638174e-02	1.258065e-01	0.000000e+00	1.756272e-01	0.000000e+00	1.212121e-01	0.000000e+00
	std	3.400728e-01	2.951234e-01	3.318989e-01	NaN	3.807302e-01	0.000000e+00	3.314340e-01	0.000000e+00
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	75%	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	max	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00
rating_bad_good_migration	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	0.000000e+00	0.000000e+00	5.645161e-02	0.000000e+00	2.389486e-02	1.000000e+00	0.000000e+00	0.000000e+00
	std	0.000000e+00	0.000000e+00	2.309781e-01	NaN	1.528129e-01	0.000000e+00	0.000000e+00	0.000000e+00
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00
	max	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00
rating_good_bad_migration	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	0.000000e+00	0.000000e+00	1.935484e-02	0.000000e+00	3.345281e-02	0.000000e+00	3.030303e-02	1.000000e+00
	std	0.000000e+00	0.000000e+00	1.378800e-01	NaN	1.799233e-01	0.000000e+00	1.740777e-01	0.000000e+00
	min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
	25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
	50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
	75%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
	max	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
rating_good_good_migration	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	1.334138e-01	9.036183e-01	7.983871e-01	1.000000e+00	7.670251e-01	0.000000e+00	8.484848e-01	0.000000e+00
	std	3.400728e-01	2.951234e-01	4.015286e-01	NaN	4.229791e-01	0.000000e+00	3.641095e-01	0.000000e+00
	min	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00
	50%	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00
	75%	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00
	max	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00
rwa_growth	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	-1.719641e-02	-8.688589e-03	1.634982e+00	-8.889946e-02	-1.490946e+00	-6.011047e-02	-1.014139e-01	1.552759e-02
	std	2.033122e-01	1.252236e-01	5.586831e-01	NaN	4.135980e-01	2.862210e-01	3.276536e-01	2.621545e-01
	min	-1.234970e+00	-1.461414e+00	6.466251e-01	-8.889946e-02	-7.091768e+00	-1.480095e+00	-1.811440e+00	-1.441692e+00
	25%	-1.703127e-02	-1.632971e-02	1.323543e+00	-8.889946e-02	-1.712464e+00	-8.392166e-02	-1.289491e-01	0.000000e+00
	50%	0.000000e+00	0.000000e+00	1.558344e+00	-8.889946e-02	-1.490351e+00	0.000000e+00	-8.315608e-03	0.000000e+00
	75%	3.478092e-03	0.000000e+00	1.809494e+00	-8.889946e-02	-1.283053e+00	0.000000e+00	0.000000e+00	5.218258e-02
	max	1.200756e+00	2.123463e+00	6.520162e+00	-8.889946e-02	6.236095e-02	1.322884e+00	1.309434e-01	1.450098e+00
tot_allocated_limit	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	1.159617e+07	3.478706e+07	2.753049e+07	1.002000e+11	7.419307e+06	7.695526e+06	2.603657e+08	5.727060e+06
	std	4.587212e+07	3.287235e+08	6.165569e+07	NaN	4.018367e+07	3.538179e+07	3.825308e+08	1.739636e+07
	min	0.000000e+00	0.000000e+00	0.000000e+00	1.002000e+11	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	3.333334e+04	5.577842e+05	1.002000e+11	0.000000e+00	1.000000e+04	1.638370e+05	1.000000e+04
	50%	5.000000e+03	1.000000e+06	5.000000e+06	1.002000e+11	9.100000e-01	2.000000e+05	4.000000e+07	2.000000e+05
	75%	1.983673e+06	2.124558e+07	2.735894e+07	1.002000e+11	2.000000e+06	1.788695e+06	4.347640e+08	2.335754e+06
	max	8.916667e+08	3.865231e+10	6.316271e+08	1.002000e+11	7.000012e+08	5.259501e+08	1.244000e+09	2.434000e+08
tot_outstanding	count	3.313000e+03	1.586400e+04	6.200000e+02	1.000000e+00	8.370000e+02	7.390000e+02	3.300000e+01	9.240000e+02
	mean	7.811694e+06	1.610217e+07	1.047210e+07	7.071330e+10	5.260665e+01	3.986586e+06	8.241758e+07	3.813956e+06
	std	3.215838e+07	1.777627e+08	2.523450e+07	NaN	5.689523e+02	1.959443e+07	1.541250e+08	1.462489e+07
	min	0.000000e+00	0.000000e+00	1.000000e-01	7.071330e+10	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	25%	0.000000e+00	0.000000e+00	8.414131e+04	7.071330e+10	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
	50%	6.358300e+02	3.000000e+04	1.085295e+06	7.071330e+10	0.000000e+00	0.000000e+00	2.085000e+06	3.085000e+00
	75%	6.821240e+05	7.718994e+06	8.077246e+06	7.071330e+10	0.000000e+00	2.093170e+05	8.135847e+07	1.915747e+05
	max	8.028967e+08	2.127293e+10	2.029433e+08	7.071330e+10	1.209244e+04	3.568445e+08	5.518288e+08	1.996681e+08

Figure C.10: Continuation of table of features statistics for the PCA dataset using K-Means

C.1.3. DBSCAN: UNCORRELATED DATASET

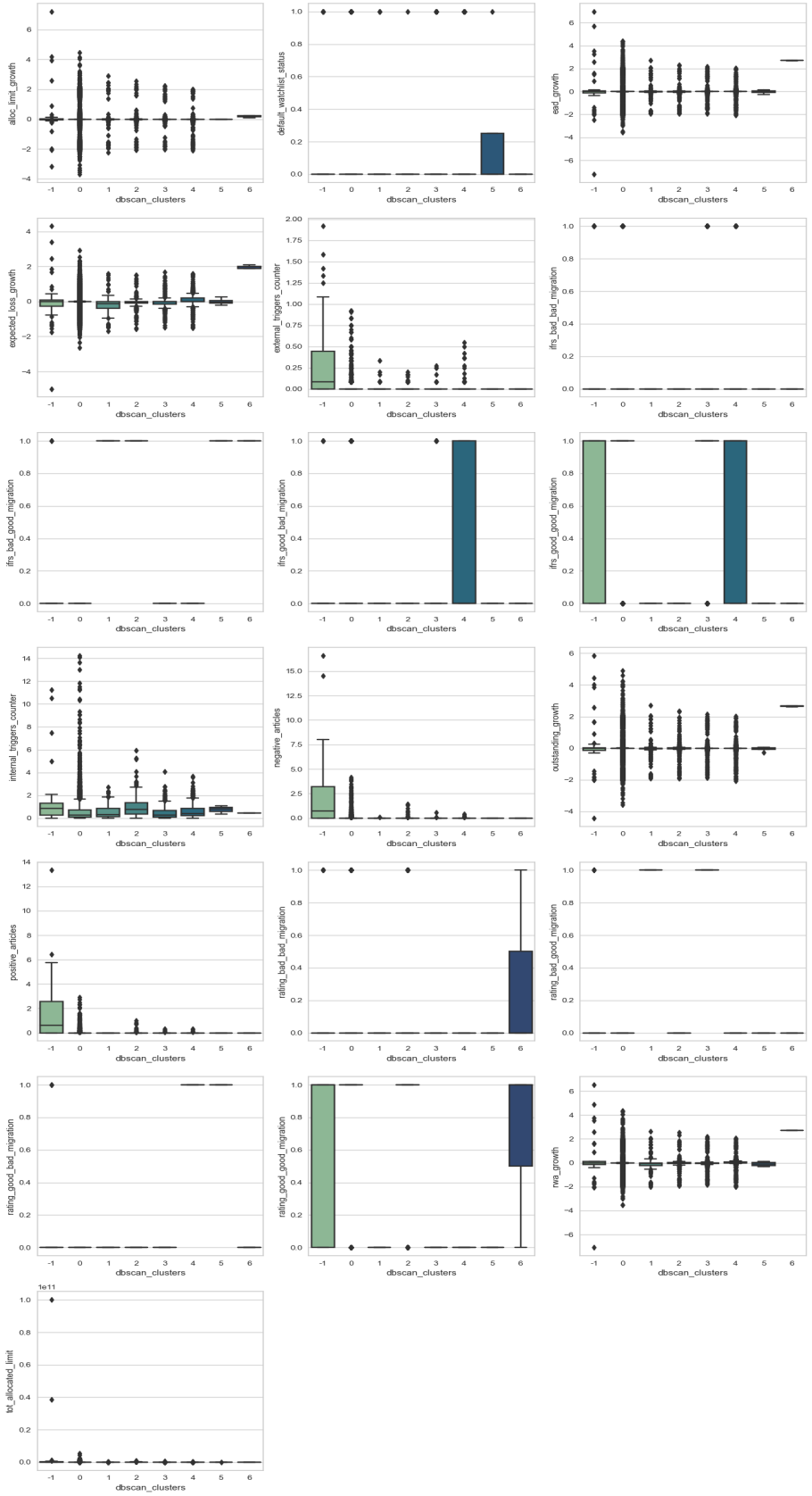


Figure C.11: Box plots of the features' values distribution for the uncorrelated dataset using DBSCAN

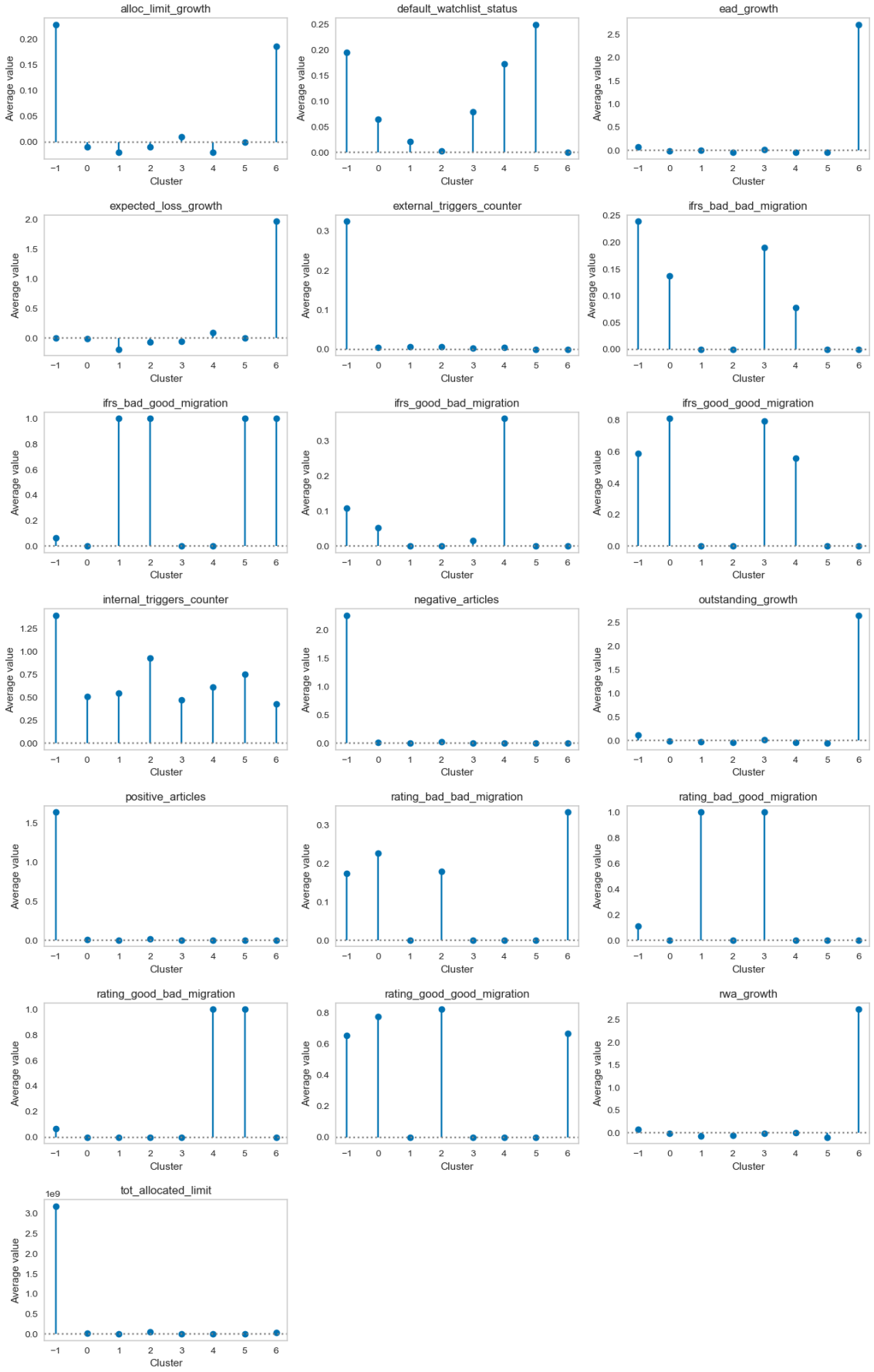


Figure C.12: Stem plots of the features' average values for the uncorrelated dataset using DBSCAN

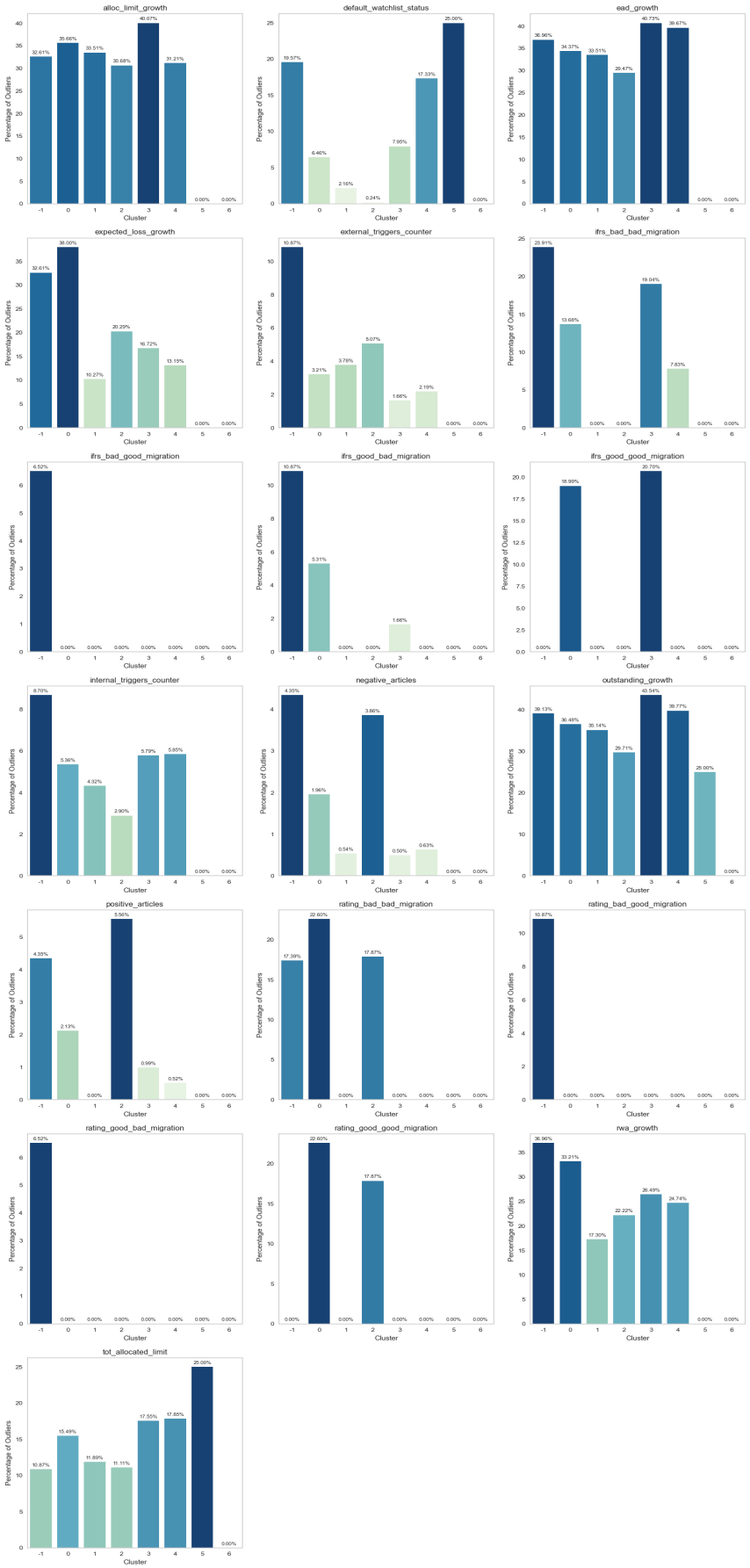


Figure C.13: Histogram of the features' outliers percentage for the uncorrelated dataset using DBSCAN

C.1.4. DBSCAN: PCA DATASET

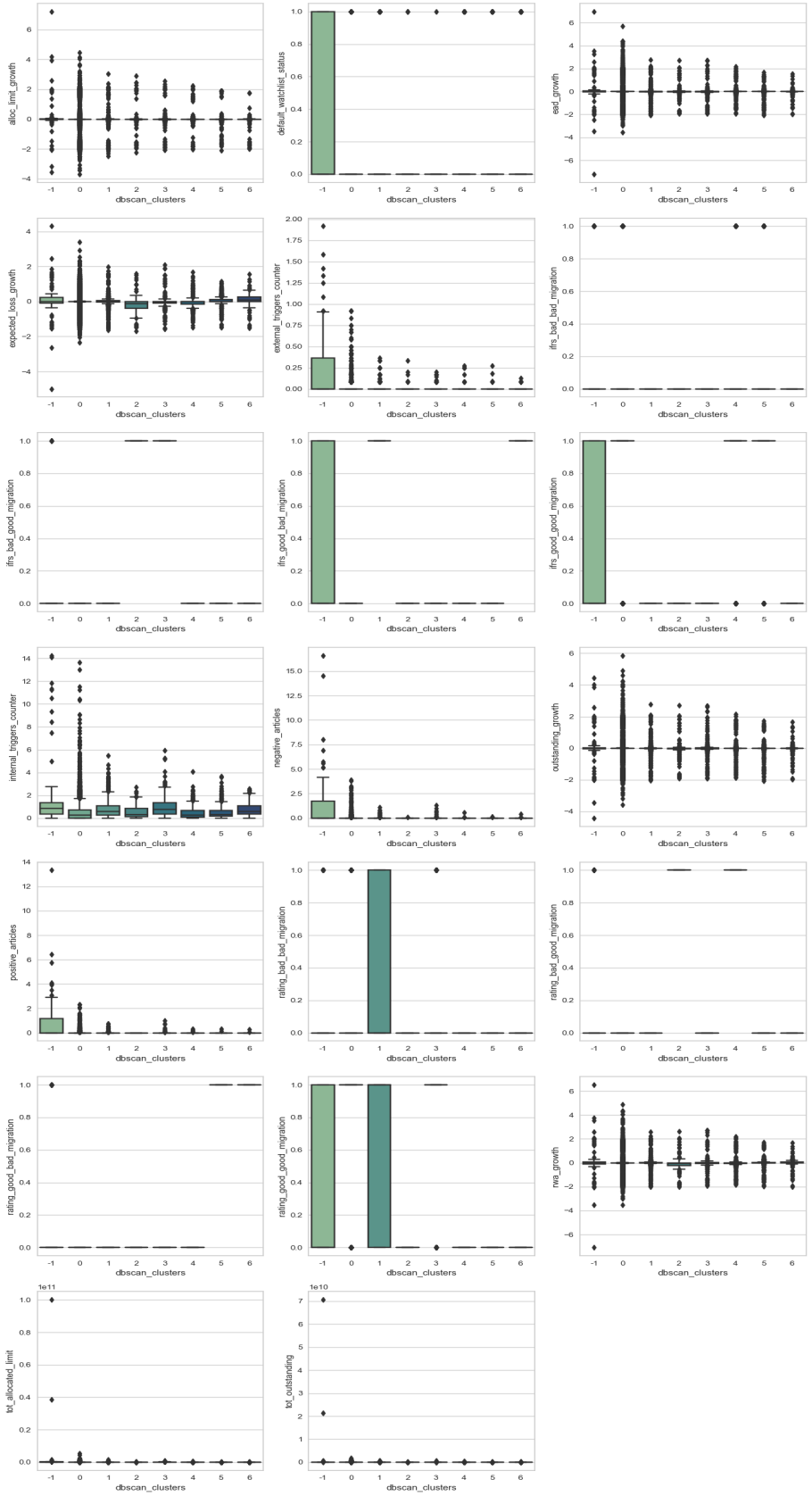


Figure C.14: Box plots of the features' values distribution for the PCA dataset using DBSCAN

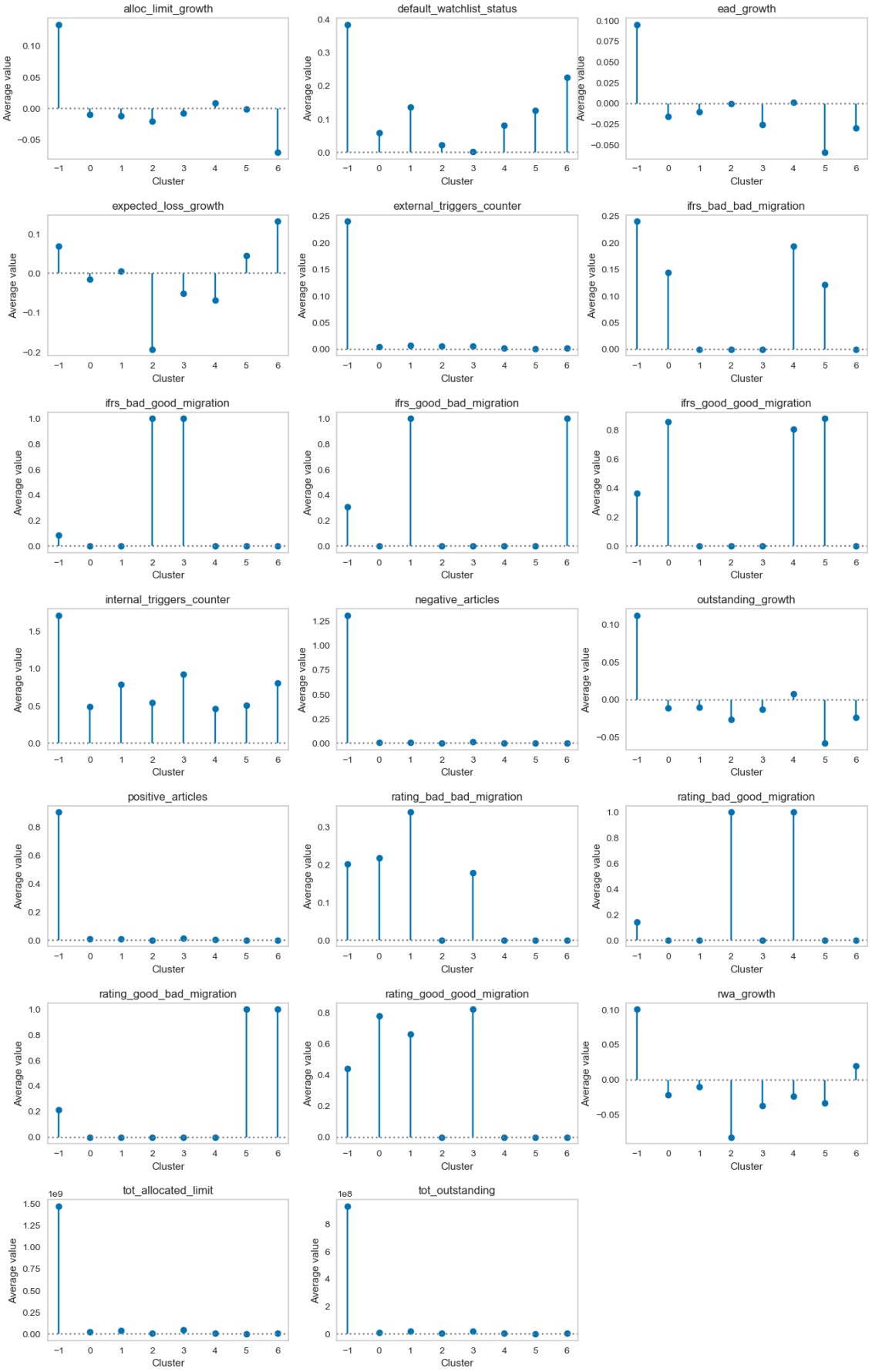


Figure C.15: Stem plots of the features' average values for the PCA dataset using DBSCAN

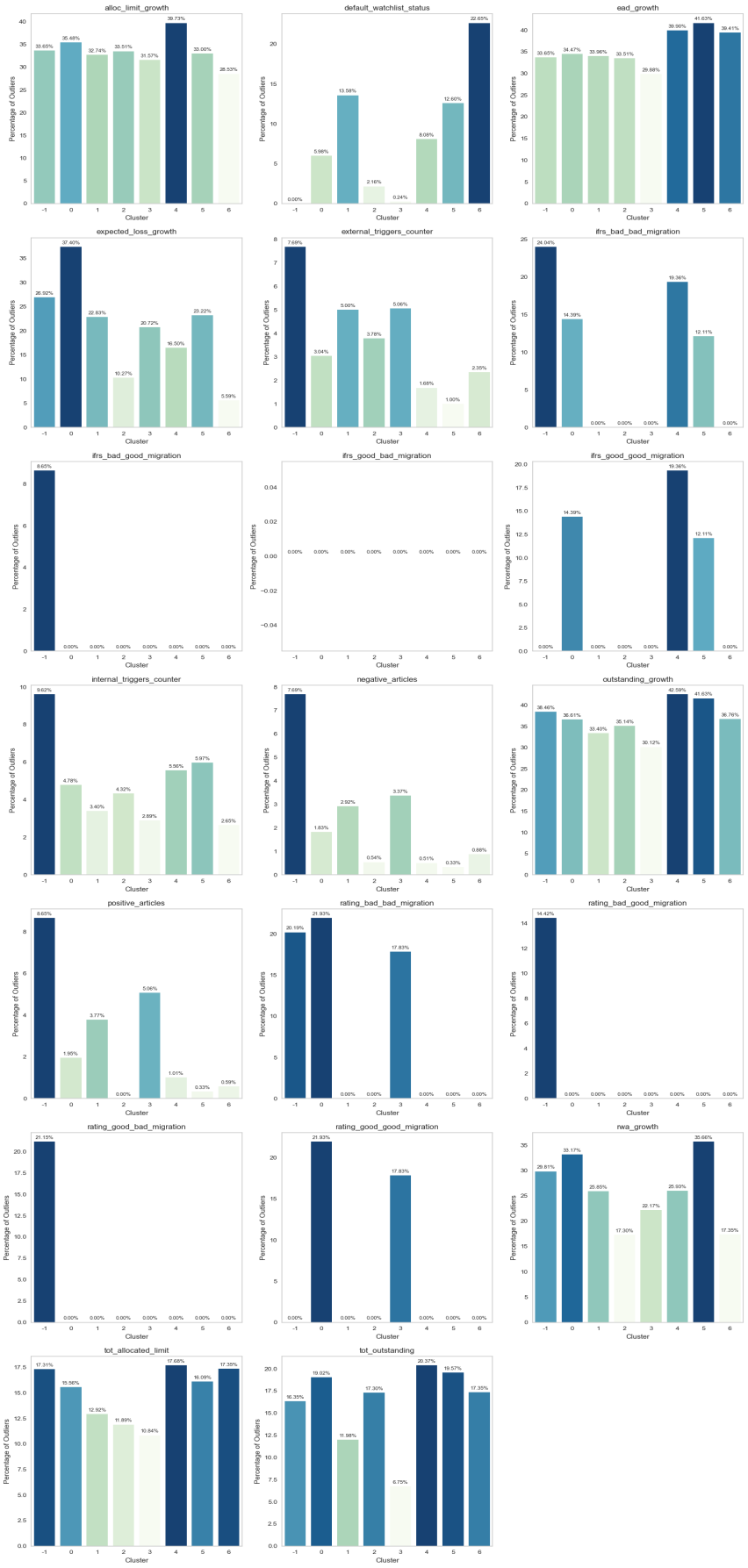


Figure C.16: Histogram of the features' outliers percentage for the PCA dataset using DBSCAN

C.2. RISK-REWARD ANALYSIS

C.2.1. UNCORRELATED DATASET: DBSCAN

UNCORRELATED DATASET: DBSCAN

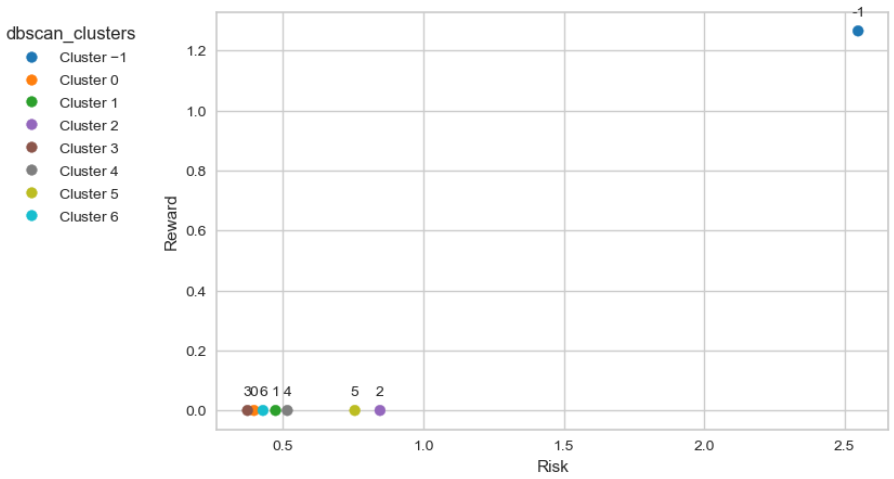


Figure C.17: Risk-Reward analysis for clusters generated from the implementation of DBSCAN on the uncorrelated dataset

C.2.2. PCA DATASET: DBSCAN

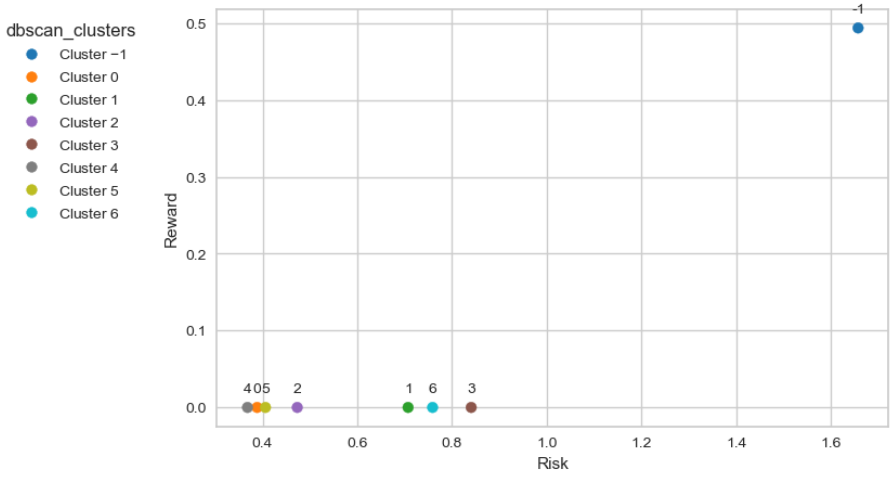
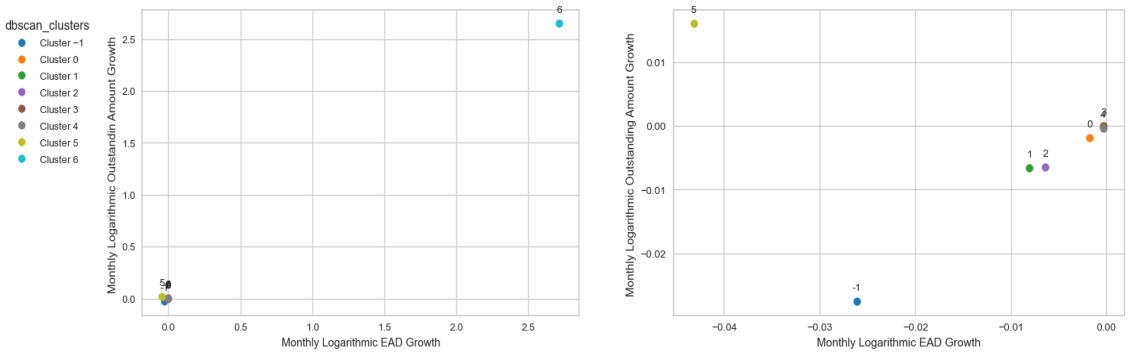


Figure C.18: Risk-Reward analysis for clusters generated from the implementation of DBSCAN on the PCA-trasnformed dataset

C.3. RISK EXPOSURE ANALYSIS

C.3.1. UNCORRELATED DATASET: DBSCAN



(a) Overview (b) Close up

Figure C.19: Risk exposure analysis for clusters generated from the implementation of DBSCAN on the uncorrelated dataset

Table C.1: Summary of the DBSCAN clusters' average Outstanding Amount growth, EAD growth, EAD-Outstanding Ratio and Total Outstanding Amount for the uncorrelated dataset

K-Means Cluster	**EAD Growth**	**Outstanding Growth**	**EAD-Outstanding Ratio**
-1	-0.026130	-0.027531	0.95
0	-0.001787	-0.001902	0.94
1	-0.008090	-0.006572	1.23
2	-0.006395	-0.006515	0.98
3	-0.000309	-0.000041	7.54
4	-0.000318	-0.000360	0.88
5	-0.043170	0.016023	-2.69
6	2.714863	2.648618	1.02

C.3.2. PCA DATASET: DBSCAN

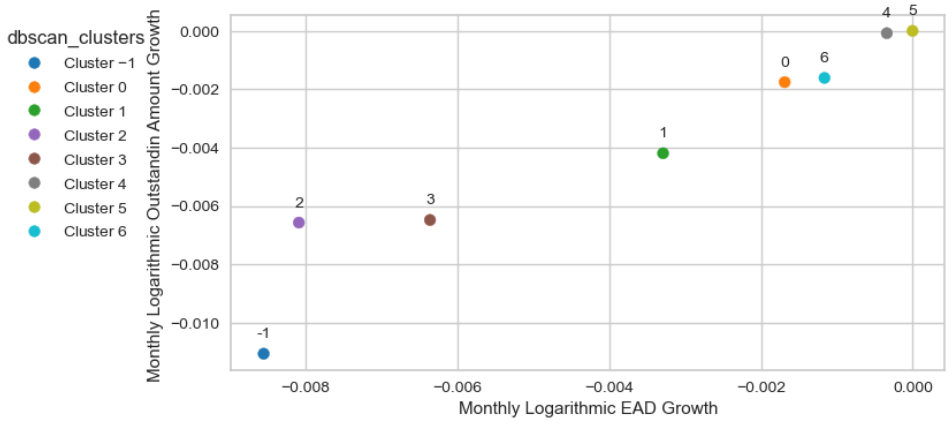


Figure C.20: Risk exposure analysis for clusters generated from the implementation of DBSCAN on the PCA dataset

Table C.2: Summary of the DBSCAN clusters' average Outstanding growth, EAD growth, EAD-Outstanding ratio and Outstanding Amount for the PCA-transformed dataset

K-Means Cluster	**EAD Growth**	**Outstanding Growth**	**EAD-Outstanding Ratio**	**Total Outstanding Amount**
-1	-0.008554	-0.011067	0.78	8.863113e+06
0	-0.001687	-0.001756	0.96	8.160464e+05
1	-0.003290	-0.004195	0.78	5.398653e+06
2	-0.008090	-0.006572	1.23	7.303806e+05
3	-0.006362	-0.006483	0.98	1.327050e+07
4	-0.000341	-0.000079	4.31	2.040366e+04
5	-0.000002	0.000000	-	6.370708e+03
6	-0.001162	-0.001611	0.72	9.362160e+05

D

Appendix D: SHAP Analysis

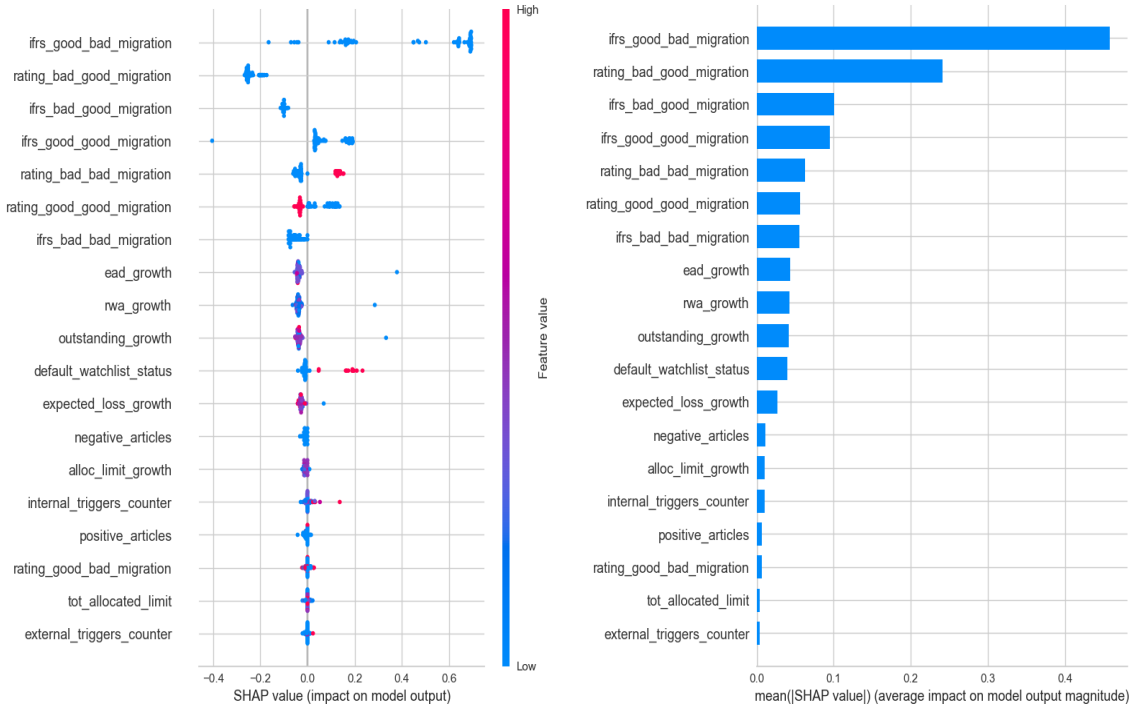


Figure D.1: Summary plot and bar plot reporting the most significant features for Cluster 1 generated with K-Means and the uncorrelated dataset

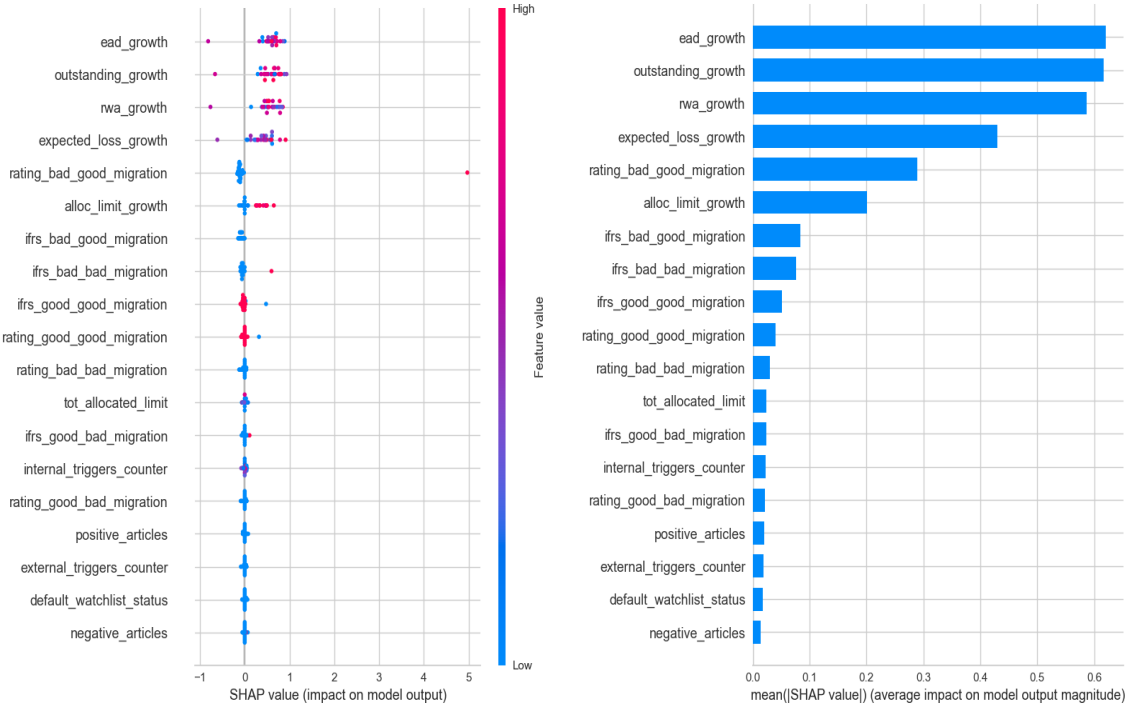


Figure D.2: Summary plot and bar plot reporting the most significant features for Cluster 3 generated with K-Means and the uncorrelated dataset

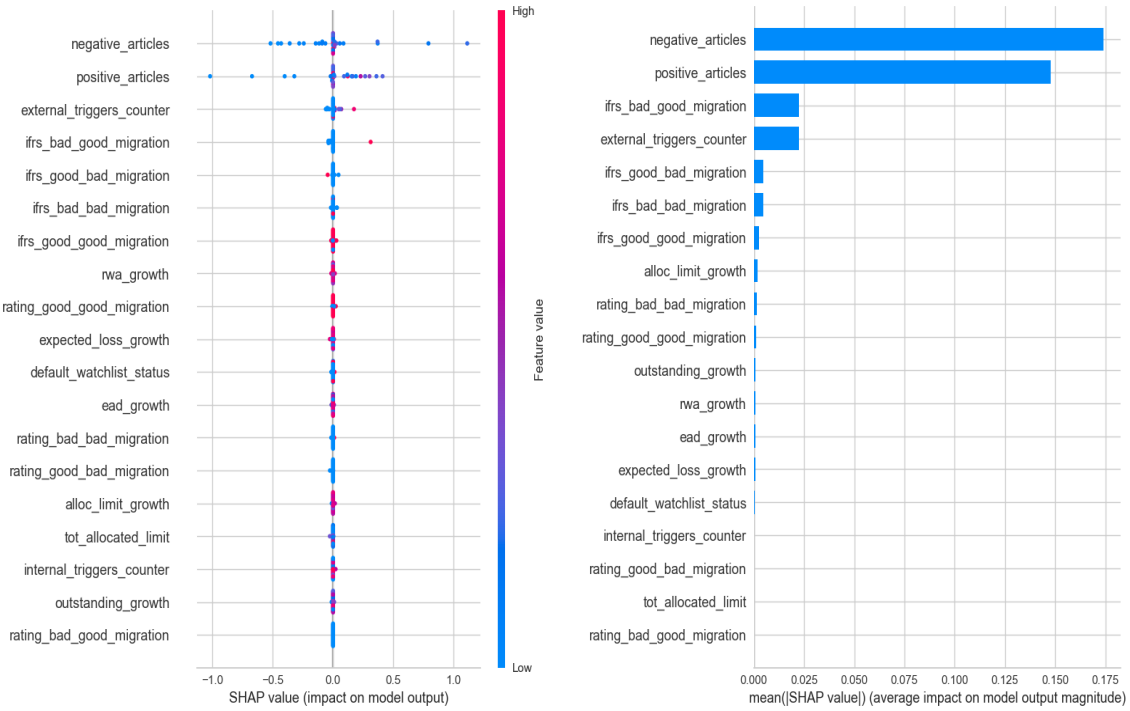


Figure D.3: Summary plot and bar plot reporting the most significant features for Cluster 6 generated with K-Means and the uncorrelated dataset