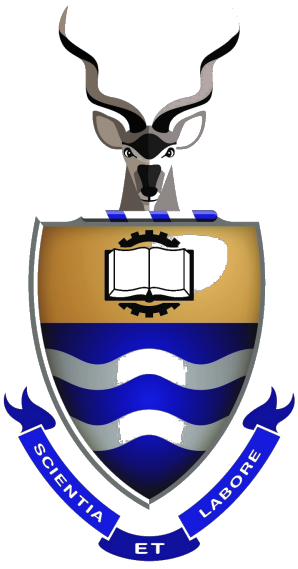
Comparative Studies on The Application of Principal Component Analysis and Recency, Frequency and Monetary Analysis to Clustering Algorithms for customer segmentation

---

Johannes Tebalo Kokozela (714188)

Supervisor(s):
Dr. Martins Arasomwan



A research proposal submitted in partial fulfillment of the requirements for the degree of M.Sc. in e-Science

in the

School of Applied Mathematics and Computer Science
University of the Witwatersrand, Johannesburg

4 September 2022

Declaration

I, Johannes Tebalo Kokozela (714188), declare that this proposal is my own, unaided work. It is being submitted for the degree of M.Sc. in e-Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

Signature

Johannes Tebalo Kokozela (714188)

4 September 2022

Abstract

Recent growth in the amount of data that companies can acquire has resulted in popularity of algorithms that assist with better understanding of the customer. Clustering has played an important role in managing of customer relationships. With the huge datasets coming from many organisations, researchers can find it time consuming to find the clusters using memory expensive algorithms. One common work-around to this problem is the use RFM (recency, frequency and monetary) related variables. In this study we propose the use of dimensionality reduction algorithm, before the application of clustering algorithm as a method to not limit ourself to transactional data. In this paper we will use principal component analysis as our dimensionality reduction method. We will then use the reduced dataset on the following clustering algorithms k-means and k-medoids clustering.

Acknowledgements

I would like to thank my supervisor Dr. Martins Arasomwan for being a great motivation and support for the undertaking of my research. I would also like to thank my wife for her understanding and Absa Bank for the funding of this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There has rapid rise in data sizes together with companies finding ways to understand both structured and unstructured data and finding new ways to store huge data sources in systems like Amazon Wed Services (AWS) and Microsoft Azure. This together with the growing competitive nature of business in each sector has led companies to trying to find different insights and understand their customers more so that they can improve the relationship and better service customers. As stated by [**?**] the biggest downfall of a company is "Companies choosing to avoid learning their customers"

The solution to better understanding of customers is to group customers into distinct segments based on customer value, customer lifetime and customer needs and want. The process of grouping customers into homogeneous groups is called customer segmentation (also known as market segmentation) [**?**]. These groups of customers are called clusters and these clusters have customers that are similar in each cluster and different in the between clusters.

Segmentation methods are usually dependent on the chosen variables in order to create clusters that are fit for purpose, as the choice of these variables are dependent on expect knowledge on the field. With several researchers coming into the field with limited expect knowledge the common choice of customer segmentation is based recency of the transactions, monetary value of the customers transactions and the frequency of transactions (RFM). This method allows the researcher to bypass the dimensionality reduction phase and feature selection by assuming that variables that speak to recency, monetary and frequency (RFM) are sufficient and might be a good substitute for expect knowledge in the field. However, as stated

by [**?**] RFM analysis tends put more weight on transactional information and less weight on the clients different information. Another approach authors take is to assume no expect knowledge in the field and rather apply dimensionality reduction on the data, which is the process of projecting data onto a lower dimensional space while retaining the variation in the data.

In this study we will compare both methods, we will be assuming no expect knowledge on the data and we will apply both RFM and dimensionality reduction to the data and apply both to a clustering algorithm.

1.1 Literature review

Research on customer segmentation primarily focused on a number of fields; Firstly is comparative study that apply different clustering algorithms to the same data to determine which one performs better. Secondly, focus on improving the performance of k-means by applying different feature selection algorithms before clustering and comparing the performance. The last group of algorithms focuses on applications of customer clustering on a new dataset or sector. Below we outline some studies that focus on the use of RFM and PCA before or after customer segmentation.

1.1.1 Customer Segmentation

In their study [**?**] proposes the novel way of selecting the initial centroids for k-means and compares the resulting clusters with the clusters obtained using fuzzy c-means and the traditional k-means algorithm. In their study [**?**] use fuzzy clustering ensemble algorithm created in the study [**?**] to create customer segmentations.

1.1.2 Clustering Algorithms

The authors use variables that speak to RFM on purchase data obtained from University of California Irwin (UCI) repository. [**?**] did an in-depth analysis of the

popular clustering analysis from different view points, including the use of RMF another in other to by-pass rigorous feature selection.
[**?**] also uses the RFM method as an input to their study in comparing k-means and k-medoids, both algorithms were compared using Davids Boulders Index (BDI) and intra-AWS distance. To obtain improved clustering result [**?**] also applied RFM to their study, where the aim of the study was to provide a method for building a clustering algorithm. [**?**] reviews the definition, application, disadvantages and advantages of using RFM model. The study further provides methods that can be used to apply the RFM method to problems.

In their study [**?**] used two datasets to compare RFM, Chi Square Automatic Interaction Detector (CHAID) and logistic regression as analytical methods for direct marketing segmentation and they found that CHAID tended outperform RFM.

1.1.3 Feature Selection

[**?**] created a two-step customer segmentation method, that uses genetic algorithm as feature selection method and then applied fuzzy clustering ensemble to create customer segments. [**?**] performed PCA as a pre-proccesing method and compared k-means and fuzzy c means clustering. They found that fuzzy C means outperformed k-means. [**?**] applied PCA and k-means in reverse on luxury goods company data, such that they started with segmenting the data with k-means, then applied PCA to the individual clusters in order be able to interpret the clusters in a lower space. Instead of segmenting clients [**?**] used both PCA and k-means to segment products by how well they're received by clients, thus how frequent they are sold. [**?**] improves the performance of k-means clustering algorithm by using self organising maps as a feature selection method.

1.1.4 Conclusion

In the studies above we observe that a lot of research in customer segmentation work on the assumption that either RFM or PCA (dimensionality reduction) are suitable for research and only a limited number of studies compare them before

application. This is the main motivation for this research and will be the primary focus of this study.

1.2 Research Question

In this study we will be investigating if customer segmentation requires that we do an in-depth feature selection phase (using dimensionality reduction) or the use of RFM as feature selection method. Thus we pose the following research questions:

- Does PCA help select the most accurate features compared to that produced using RFM analysis?
- Between k-means and k-medoids which algorithm produces better clusters[1]?
- Between principal component analysis and recency, frequency and monetary value analysis which of the two feature selection algorithm aid in producing clusters that are easy to interpret?
- which of the the two feature selection algorithm i.e. PCA and RFM is more accurate?

1.3 Research Aims and Objectives

In the previous sections we outlined the literature on the subject of customer segmentation and posed our research question, we are now ready to state the research aim and objectives that we would like to achieve to complete this study.

1.3.1 Research Aims

The aim of this study is to compare component principal analysis against recency, frequency and monetary analysis on clustering algorithms for customer segmentation.

1.3.2 Objectives

1. Compare the effect of PCA and RFM on k-means and k-medoids.
2. Compare the performance of k-means and k-medoids using BDI.

[1] Better in this study is defined by the Davies Bouldin Index (BDI) and the ease of interpretation by visualisation.

3. Determine which algorithm produces results that easier to interpret.

4. Determine the accuracy of PCA against that of RFM.

1.4 Limitations

In this study we will be using data that is specific to a certain market and thus this will might limit the results to that market. Furthermore, the algorithms in this study could be applied in a different way or improved for a better performance, but here we will limit ourself to their basic form. As an example k-means algorithm can use a variety of distance measure, but here we will limit our study to the euclidean distance.

1.5 Overview

In chapter 1 we briefly introduced the subject of our study which is customer segmentation, and then we gave a brief overview of the literature that preceded this study. In chapter 2 we will provide details of the methods that we will undertake in the study and finally in chapter **??** we will provide details of the work and the timelines that we plan to follow.

Chapter 2

Research Methodology

In this chapter we will provide details of algorithms that we will be implementing for our proposed study. We have chosen to develop k-means and k-medoids as our clustering algorithm, this is primarily motivated by the popularity of the algorithms in the studies that we looked at. There other clustering algorithms, such as hierarchical clustering and self-organising maps. Furthermore, the are a number of feature selection methods that could be applied to this or similar problem, one limitation they mostly have is the depends on the response variable, given that our problem is unsupervised, we're limited to dimensionality reduction and expect knowledge.

2.1 Research design

In the rest of this section we will be referring to Figure 2.1, we will give a brief explaining of the figure here. The figure should be read in a top down approach. We begin with the sourcing stage where we will be collecting data and then do some preprocessing, this would have to be transactional data so that we can apply RFM to the data. The next phase is feature selection and we will be applying the methods to answer our posed research questions. The next phase will be our model development phase, the models we will be implementing are k-means and k-medoids. We will then evaluate the model performance using DBI.

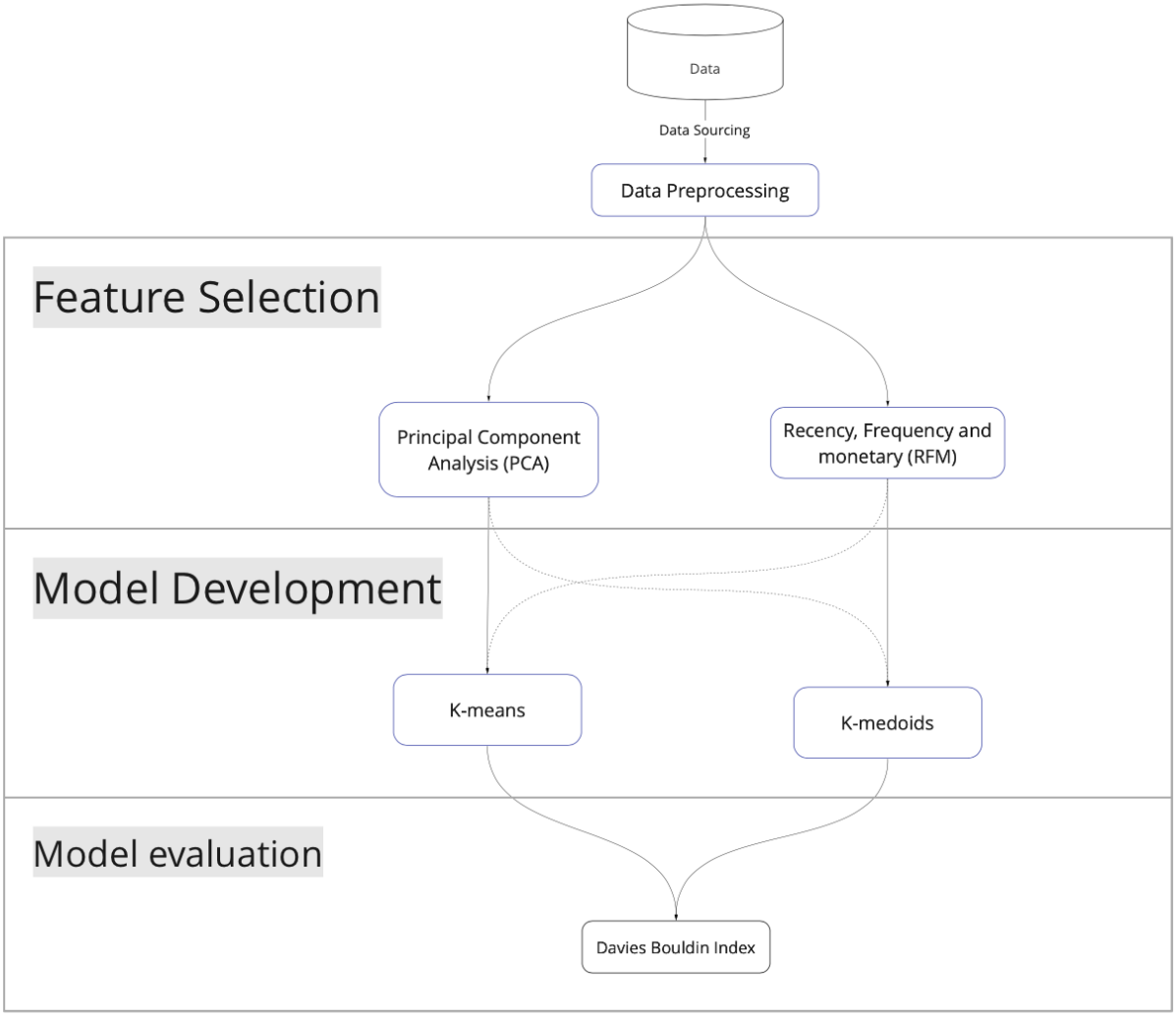


FIGURE 2.1: Methodology flow-chart

2.2 Methods

Until now we have discussed the algorithms that we will be using in out study but have not outlined the details of the study. In the subsections that follow we will be give a brief outline of RFM, PCA, K-means and k-medoids, this will be a brief outline without going into the details of the mathematics.

2.2.1 Recency, Frequency and Monetary Analysis

- Recency - The difference between the transaction date and the maximum date of the data span.
- Frequency - The the total number of transaction a customer has made in the span of the dataset.
- Monetary - The total amount a customer has spent in the span of the collected dataset.

2.2.2 Principal Component Analysis (PCA)

Here we outline the steps that we take to perform PCA

1. Calculate the covariance matrix thus creating an adjusted data.
2. We calculate the eigenvalues and eigenvectors in the adjusted data.
3. Choose components and form a feature space (reduced data space).

2.2.3 K-means and K-medoids

We will outline both k-means and k-medoids together.

1. We begin by choosing the value of k (the number of clusters) - we will use the elbow method for both algorithms.
2. We then assign centroids to random value for k-means and for k-medoids we will randomly assign centroids to data-points.
3. We will constantly update the centroids according to the euclidean distance.
4. The stoping condition will be: If there is no change in the centroids values.

2.3 Ethical Considerations

The ethics has to be accounted in the application of this model. Segmentation models are discriminating in nature and these can be used in a bad way. Therefore, we suggest that when working with this model to adhere to the model regulations. Furthermore, we will try to remove variables that are related to race and location (zip codes) as these tend to create a model that can be seen as biased towards certain demographics.

Chapter 3

Schedule of Work

We will be presenting the work schedule that will be following to achieve the objectives set in chapter 2. Due to the limited amount of time available our schedule be aggressive. These times are subject to changes should there be any changes should there be any changes in the work we will undertake.

3.1 Schedule of Work

Figure **??** is the graphical representation of the scheduled plan. The first row of the figure is the year and third row is the month number of the year. From figure **??** we see that the plan is to complete the study by the 10th month (October) of the year. We expect to take three months on data sourcing and preprocessing thus the expect end data will be by the end of July. We expect to then begin model development in July and finish on the month of October thus begin work on drafting the final report.

The final phase in our study will be to write up the report documentation of the results we will have gathered. Then finally we gather all the results in a poster that will be presented to an audience.

3.2 Potential Difficulties

The most predictable difficulty in a modelling problem is the data, sourcing data can be a very difficult task as companies have very strict policies on sharing data about customers. Another data problem is the fact that data preprocessing requires

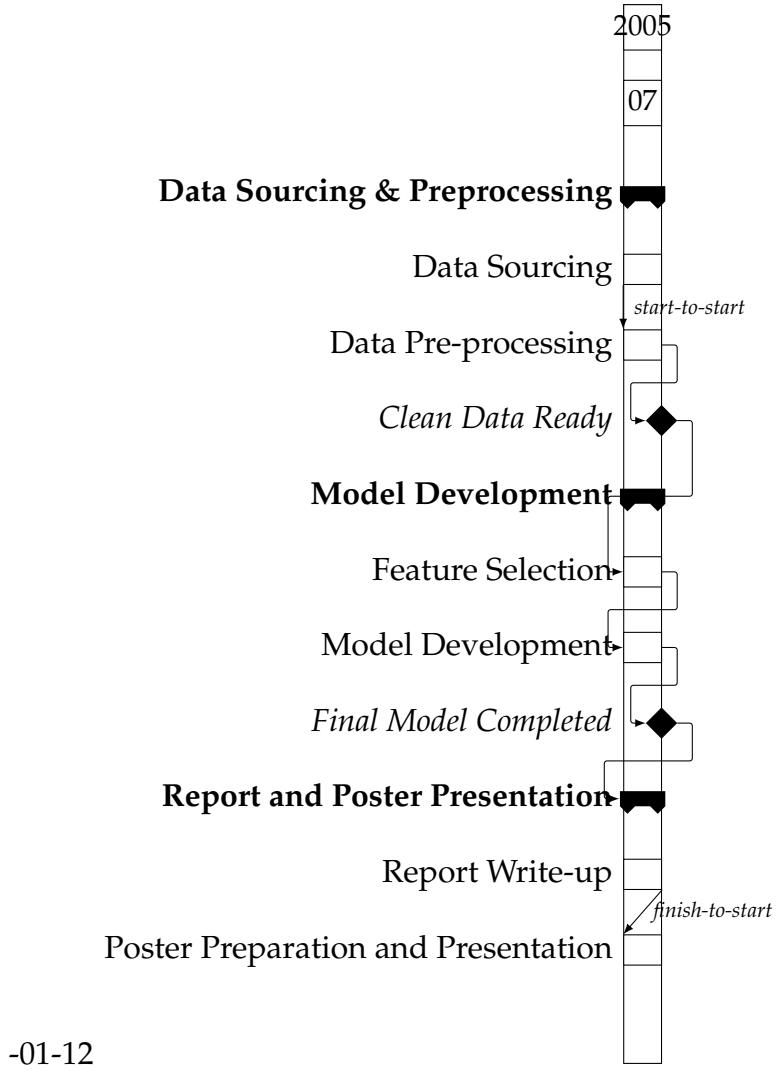


FIGURE 3.1: Scheduled Plan

some domain knowledge of the field in which the data comes from. The final predictable difficulty is the difficulty that comes with interpretation of dimensionality reduction as this is a different space to the original data.

Chapter 4

Conclusion

To conclude our study we reiterate the gap that we identified in the current literature, which was the fact that studies in the past tend to choose between domain knowledge based customer segmentation in RFM or focus more on dimensionality reduction but not both. Our aim is to determine the if transactional data alone is sufficient to create accurate customer segments. The results of this study could contribute to customer segmentation by either asserting that transactional data alone is sufficient or there might be a need (in some cases) to include other datasets like demographics.