UNIVERSITY OF NAIROBI

MASTER OF SCIENCE IN COMPUTATIONAL INTELLIGENCE

A CLUSTERING APPROACH TO MARKET SEGMENTATION USING INTEGRATED BUSINESS DATA

ISABEL MAKARA

P52/34108/2019

Supervisor

Dr. LAWRENCE MUCHEMI

A Project Proposal Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computational Intelligence, School of Computing and Informatics, University of Nairobi.

September 2020

DECLARATION

This project, as presented in this report, is my original work and has not been presented foraward in any other university.

Name: Isabel Makara **RegNo:** P52/34108/201

Signature:_____ **Date:** 20/9/2021

This project has been submitted with my approval as the University supervisor.

Name: Dr. Lawrence Muchemi

Signature:_____ **Date:** 1st Sept 2021

DEDICATION

I would like to dedicate this work to my mother, Virginia Wangari. You have been a force to reckon with, the wind beneath my wings. Thank you for the opportunity to do this Masters course and for all the sacrifices you have made for me. I love you immensely.

I would also like to dedicate this paper to my fiancé. Thank you for your endless support and encouragement all through this journey. We have finally made it.

ACKNOWLEDGEMENT

I would like express my sincere gratitude to my heavenly Father, my God, who has brought me this far. And His ever sufficient grace which has carried me through this course till the end. Thank you.

I am also very grateful to my supervisor, Dr. Lawrence Muchemi for all the guidance and assistance provided throughout this project.

My family, my number one fans and cheerleaders. Thank you for all your support throughout this journey.

My fiancé, who has held my hand throughout this Masters journey, thank you for your endless support and encouragement.

To my friends and colleagues who walked with me throughout this journey, giving me words of encouragement. I am eternally grateful.

To the school, lecturers and the supporting staff, I appreciate this opportunity accorded to me, thank you for your support.

Definition Of Important Terms

Clustering algorithm: An unsupervised learning method used to draw references from datasets without the guidance of previously defined references or classes

Market segmentation: The process of sub-dividing a heterogenous market into smaller homogenous segments.

Small businesses: A business with between 10 and 50 employees with an annual revenue turnover of more than Ksh 10 million but below Ksh 50M.

Ground truth: Information or data used to check the results of a machine learning algorithm based on real world labels

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Market segmentation approaches applied by small businesses in Kenya have mostly been based on very limited customer factors. This includes geographic, demographic, behavioral, and psychographic characteristics of the customer. These approaches have not entirely brought out the nature of the customers in the business. In some cases, the approaches have been based on incorrect assumptions and have also led to challenges like potentially ignoring new markets and difficulty in keeping up with changing customer needs. This research seeks to use a clustering algorithm to carry out market segmentation based on more variated and integrated data that could give more information on customer habits. The dataset that was used in this research contained integrated data on various customer and business facets. The research used the spectral co-clustering algorithm to bring out various traits on the business customers that could then be used to segment the customers into more effective markets. This research eventually brought out different market segments with varying characteristics all based on the best features in the integrated business data.

This study showed that there is more data available in a business that can be used for market segmentation other than just geographical, behavioral, demographic and psychographic data. It also brought out the importance of feature selection in a dataset since different features may have different effects on sales on a business and the overall performance of a market segment.

This study also contributes to research by identifying other features other than geographical, demographic, psychographic and behavioral factors that could be used to identify market segments in a small retail business. This approach is also more effective since the use of a clustering algorithm enables the discovery of patterns in business data that would not have been easy to spot with the naked eye, and the use of up-to-date business data aids the business in keeping up with customer habit changes and prevents missing out on potential markets.

CHAPTER 1: INTRODUCTION

1.1 Background

The performance of 68.1% of small businesses are affected to a large extent by the marketing strategies they adopt (Muola, 2017). By selecting the right marketing strategy for its products, a company can manage to correctly utilize its resources and increase its profitability. (Muhammad Adi Khairul Anshary*, 2016) Companies in Kenya have recognized that they cannot appeal to all consumers in the same way since consumers are widely scattered in their needs and practices. (Njoroge, 2015) This has brought about the need to segment their markets based on various characteristics to enable them to reach their customers more effectively.

The market segmentation approaches used have mostly focused on the geographic, demographic, behavioral, and psychographic characteristics of the customer. These approaches have been fixed to only customer data on geographical location or demographic information (gender, age, etc) or lifestyle and personality characteristics. The consideration of one facet or feature of customer data to create market segments for a business is not enough. This is because the other features and factors held in business and customer data hold more insights which could create more elaborate and effective customer segments when they are all considered during customer segmentation. This research aims at carrying out a multi-faceted approach to market segmentation where we shall use various data sources and data features together to come up with customer segments for small retail businesses.

Chloride Exide company carried out market segmentation based on customer behavior data like user status, user rates, benefits, and geographical location. These factors, though important, were not enough for the company to create concise market segments bringing about the challenge of ignoring potential market segments and not managing to keep up with the dynamic changes to their customers' needs. The company however highlighted the following aspects to be important in influencing the customers' purchase decisions: education, gender, income levels, and use of additional services like warranties and discounts. These factors if included during the market segmentation process could lead to a better view of the market clusters for the company. (Yabs, 2014)

Kenya Commercial Bank in Mombasa carried out market segmentation through the geographical, demographic and behavioral segmentation approaches. In the end, they identified geographical segmentation as a very effective approach. Demographic and behavioral

segmentation was also quite significant but psychographic segmentation was not effective for the organization. (Wayabila, 2020). Considering only one of the segmentation approaches was highly effective, it was important for the organization to consider alternative approaches to segment their customer markets. The use of more data for market segmentation was highlighted as channel that would enable the bank to carry out its customer segmentation goals in a more effective manner.

The Kenya Commercial Bank in Kisii county mainly used geographical market segmentation for products that were needed by a large and diverse group of customers. This was with the assumption that the entire geographical region comprised of customers who had same behaviors towards the products, an aspect that was not entirely the case since customers in the same geographical region also had different product preferences. (Nyabwari, 2017)

It was then important for the bank to incorporate more variated data in their customer segmentation process to eliminate these assumptions and in the process come up with data driven market segments that were more effective for the organization.

Customer segmentation through clustering is the "use of a mathematical model to discover groups of similar customers based on finding the smallest variations in data". (Optimove, 2020) Clustering enables the use of a large variety of data on a given market since it has the capability to learn from diverse data sources like information from loyalty cards, point of sale systems and existing datasets to better categorize the different customers into more homogenous segments that can be easily targeted during marketing. It also addresses the challenge of needing to keep up with the changing customer needs in a company since the algorithm constantly learns from the data fed into the business' system, thus giving up to date market segments.

1.2 Problem Statement

Organizations in Kenya have been limiting their consideration to only a few factors during market segmentation. These have been either geographical or behavioral factors and, in some cases, demographic data has also been used. (Wayabila, 2020) These approaches have mostly been based on incorrect assumptions and have not been completely effective. (Nyabwari, 2017) The use of this approaches has also led to challenges like potentially ignoring new markets and difficulty in keeping up with changing customer needs.

There is thus the need to come up with a market segmentation approach that is more effective by using more variated and dynamic data factors that could give more information on customer habits.

1.3 Objectives

Overall Objective

The aim of this project was to create a machine learning model that would use a variety of sales and customer data to create effective market segments.

Research Questions

The following research questions were used to attain the overall objective of the project.

1. What are the sources of data that hold information on customer data and business sales?
2. What are the features that influence sales in a given business?
3. What is the best algorithm to use to cluster customers into effective market segments?

Research Objectives

The above research questions then advised on the following research objectives

1. Identify the sources of data that hold information on customer data and business sales
2. Identify the features that most affect sales in a given business
3. Identify an optimal algorithm that can be used to cluster customers into effective market segments
4. Develop and validate a machine learning model that will use these different sources of data to design market segments for a given business.

1.4 Justification

The success or failure of a business has been greatly attributed to the marketing strategy that the business employed (Muthee Janet, 2014). The market segmentation approaches currently used by organizations have not been completely effective due to the limited nature of data used. It has brought about challenges in bringing out the actual structure of the market thus raising the risk of ignoring potential market segments that would have been captured with the use of more variated data. It is therefore important to use a wide variety of data factors to create customer segments.

1.5 Significance

Small businesses in Kenya will benefit from this research since it will provide a more effective approach to come up with market segments. Through the use of integrated data, the small businesses will manage to consider a wider array of factors regarding their customers and the business during this process. This will help to alleviate any assumptions made during market segmentation processes.

The use of a machine learning algorithm on the integrated business data will enable the business to identify patterns in their data that could not have been possible with only one source of data. It will also help the business in keeping up with the dynamic changes in customer needs through the different results given by the algorithm at different circumstance. This will lead to the creation of more effective market segments in small businesses.

1.6 Assumptions And Limitations

a) This research will focus on small retail business in Kenya
b) There may be other factors or conditions that may affect the marketing of a given product that may not be represented by the data collected during the project.

CHAPTER 2: LITERATURE REVIEW

2.1. Introduction

Market segmentation is the process of sub-dividing a heterogenous market into smaller homogenous segments. (Hayes, 2020) It aims at identifying similar traits among customers and using them to divide the customers into groups that can have a certain similar response to a marketing action. (Tarver, 2020). Market segmentation in businesses has mostly been carried out using the following main approaches:

Geographic segmentation: This involves dividing a market based on geographical locations.

Demographic segmentation: Dividing the market based on age, gender, family size and income

Psychographic segmentation: This is segmentation based on customer's personalities, lifestyle and social class.

Behavioral segmentation: Customer segmentation based on the user response or usage statistics of a given product. (Ghose, n.d.)

Most businesses in Kenya have been carrying out market segmentation by considering either one or a combination of some of the above approaches.

Chloride Exide company carried out market segmentation based on geographical location, and behavior. Though the impact of market segmentation on their revenues was positive, the organization still faced the challenge of ignoring potential customer segments and keeping up with constant changes in the customer needs. These are challenges that could be addressed by using more data during market segmentation to ensure that the company obtained a better understanding of the market characteristics. Chloride Exide also identified other factors that had an impact on their sales like education and income levels. These latter factors had not been considered during the market segmentation process. Including them would have improved the effectiveness of the market segments. (Yabs, 2014)

The Kenya Commercial Bank in Kisii county mainly uses geographical market segmentation for products that are needed by a large and diverse group of customers. This is with the assumption that the entire geographical region comprises of customers who have same behaviors towards the product which is not usually entirely the case. Segmentation based on behavioral and socioeconomic characteristics is also highly used at the branch. The disadvantage to this is the fact that the descriptive nature of an individual did not entirely bring out their desire to buy a specific product. The assumption that members in the same demographic behave identically toward a given product is also not entirely true since different individuals in the same demographic tend to have varying tastes and traits. (Nyabwari, 2017). It is then important to identify and incorporate more data with better attributes that can be used to correctly identify market segments. The bank also found it important to consider using behavioral data like credit data, online banking and credit card transactions. These avenues provided a lot of customer data that could be used to track the customer real time needs. It was however resource intensive to carry out analysis of this data without the help of machine learning algorithms. (Nyabwari, 2017) Machine learning algorithms are resource efficient and at the same time give real time analysis on the bank's customers, enabling the development of concise and up to date market segments.

The consideration of only one or two aspects from the four mentioned approaches (demographics, geographical, psychographic and behavioral approaches) for market segmentation, is not enough to correctly categorize a customer into a given segment. There is the need to consider more data like transactional data, product data and loyalty card data among

other forms of business data to get a better understanding of a certain segment of customers. This approach requires the use of machine learning algorithms that can go through the numerous amounts of data required for the process and come up with patterns that can create effective market segments.

2.2. Existing Work

2.2.1. Tweet-Based Target Market Classification Using Ensemble Method

"Ensemble methods involve the use of multiple learning algorithms to train models with the same dataset". (Ghoshal, 2020) Classification models were constructed using two ensemble methods, bagging and boosting, on a dataset of 3000 tweets. The aim was to classify the twitter followers based on the tweets of the given account. There were different versions of the algorithm applied for performance comparison. These were bagging CART, AdaBoost CART and CART. CART (Classification and Regression Tree) was the conventional method while the other two were the ensemble methods. The accuracy of the algorithms was measured by comparing the experimental data being correct compared to the test data.

The bagging algorithm generated the highest accuracy with 85.20% accuracy. The accuracy of CART was 83.30% prior to the use of the ensemble method. The bagging CART algorithm also generated the highest value of precision and recall in product classification at 86.9% for precision (The percentage of the results which are relevant) and 85.2% for recall (The percentage of total relevant results correctly classified by the algorithm). (Muhammad Adi Khairul Anshary*, 2016) The use of ensemble methods for classification was therefore more effective that the use of conventional classification algorithms without the ensemble methods.

2.2.2. Use of K-Means on Business Data for Efficient Customer Segmentation: A Strategy for Targeted Customer Services

Clustering was carried out on a customer dataset belonging to a retail business in Nigeria for the purpose of customer segmentation. This was in order to advice on the customized marketing programs, business decision support and customer product associations for the different customer clusters. The dataset consisted of two attributes which were the average amount of goods purchased by a customer per month and the average number of customer visits per month. To come up with the end result, there were various operations that were carried out on the data.

Feature normalization was carried out to adjust the data elements to a common scale. This was in order to enhance the results of the clustering algorithm. The Z-score normalization technique was used to normalize the features between a scale of -2 to +2.

The K-Means algorithm was then run on the data set. The initial centroids were chosen using the Forgy method. Here the data points were randomly selected as cluster centroids. The other data points were then assigned clusters based on their square Euclidean norms from each centroid. The cluster assignment process was repeated until a hundred iterations when there was no more change to the cluster centroids. After this, the algorithm could now cluster the data points accurately. A purity measure was then carried out to evaluate the extent of accuracy to which each data point was clustered. It was 95%.(Pascal et al., 2015)

The K-Means algorithm gave a 95% purity measure after 100 iterations. This outcome is good, though the user is required to specify a k value for the initial number of clusters which if done wrong could give a wrong outcome. The initial centroid values are also equally sensitive thus a lot of thought needs to be put into this part of the process too to ensure the right clusters are developed. The K-Means algorithm is also very sensitive to outliers thus very thorough data preprocessing and ways to handle outliers is required. (Dr. Manju Kaushik, 2014)

2.2.3. Customer Segmentation and Profiling using Demographic Data Obtained from Daily Mobile Conversations

Multiple clustering algorithms were used for customer segmentation based on data collected through daily mobile conversations in Kenya. Four pre-existing demographic characteristics were used to aid in profiling the customers. These were gender, age group, region and living standards measures. K-Means, Partition Around Medoids and Hierarchical clustering algorithms were then compared using various validation techniques. Hierarchical clustering performed the best in Connectivity and Silhouette evaluations. A comparison of stability measures among the algorithms resulted in a tie between Hierarchical clustering and partition around medoids where hierarchical clustering had 4 clusters and partition around medoids had 12 clusters.

A combination of both internal validation and stability measures were used to create an aggregated ranking which resulted in the hierarchical clustering algorithm performing best on four of the seven measures. This brought out agglomerative hierarchical clustering as the best algorithm to use for the customer segmentation process. It was then used to cluster customers based on the data collected. (Samuel Kamande, 2018)

2.3. Clustering

Clustering is an unsupervised learning method used to draw references from datasets without the guidance of previously defined references or classes. It binds the data points present in a dataset into a number of groups with a given similarity for each group. This is done by repeatedly comparing the patterns of the input data until stable clusters are formed based on various identified similarities. (Chinedu Pascal Ezenkwu, 2015)

There are various methods used in clustering:

- "Density based clustering methods create clusters based on the density of the dataset region. They detect areas where points are concentrated and separated by empty areas. There are three different methods under the density based clustering.
    - Defined Distance(DBSCAN): Uses a defined distance to separate dense clusters from sparser noise.
    - Self-Adjusting (HDBSCAN): Uses a range of distances to separate clusters of varying densities from sparser noise.
    - Multi-scale(OPTICS): Uses the distance between neighboring features to create a reachability plot which is then used to separate clusters of varying densities from noise" (ArcGIS Pro, 2020)
- Hierarchical based clustering methods form clusters based on the hierarchy of the previously formed clusters. The approach to form the clusters can be bottom up (Agglomerative) or top down (divisive).
- Partitioning methods form K number of clusters based on an objective criterion similarity function.
- Grid based methods are based on a data space formulated into a finite number of cells that form a grid based on their densities.

From these methods there are various clustering algorithms:

K-Means clustering is a clustering algorithm where the number of classes to be used is first selected and their center points are randomly initialized. Each data point is then classified in the class whose center is closest to the data point. Once the point is classified, the group centers are recalculated by calculating the mean of the vectors in the group. This is done iteratively until there is minimal change in the group center values. K-Means is a fast algorithm since the

only computations done are the distances between points and group centers. The disadvantage is that the need to initially select classes and their centers may skew the output since it may not be entirely accurate. It also doesn't work well when the clusters are not circular. (Seif, 2018)

K-Medians clustering algorithm runs almost like K-Means algorithm but it uses the median to recalculate the group centers. It is more effective than K-Means since it is less sensitive to outliers but the iterations for computation of the center values is much slower for larger datasets. (Seif, 2018)

Since the algorithm uses the distance metric to cluster points in a dataset. As dimensions(columns) grow in a dataset, the distance to the nearest neighbor becomes undistinguishable between the centroid and various points. This is identified as the curse of dimensionality. (M. Pavithra, 2017) This factor weakens the K-Means algorithm and other distance-based clustering methods when it comes to high dimensional data.

Density-Based Spatial Clustering is a density-based algorithm which is used to detect clusters with different sizes and shapes. (Qi Xianting, 2016) It begins with a starting data point that has not yet been visited. It then extracts neighborhood points using a neighborhood threshold epsilon e. If there are a sufficient number of points in the neighborhood according to minPoints, then the current data point becomes the first point of a cluster and the e neighbor points become part of the cluster. Otherwise, the point is labelled as noise. The process is repeated for any new points that have been added to the cluster group. Once one cluster is completed, an unvisited point is the processes and the clustering process is repeated again until all points in the dataset are in a cluster or marked as noise.

The density based spatial clustering algorithm has the main advantage of not requiring a set of clusters to be pre-defined. It also defines outliers as noise and can find randomly shaped and randomly sized clusters. Its disadvantage is that it doesn't perform as well as when there are clusters of varying density since the distance threshold e and the minPoints will vary from one cluster to another. (Seif, 2018) It also faces the challenge of neighbor explosion and redundancies in merging, a factor which makes it unfit for high dimensional data. (Thapana Boonchoo, n.d.)

Gaussian Mixture Models (GMMs) is an algorithm that assumes that the data points are Gaussian distributed. It uses the mean and standard deviation to describe the shape of the clusters. Each Gaussian distribution is assigned to one cluster. The clustering process starts with selecting the number of clusters. The Gaussian distribution parameters are then randomly

initialized. The probability that each data point belongs to a given cluster is calculated based on how close it is to the Gaussian center. (Seif, 2018)

2.4. Gap

Most small businesses in Kenya have been using Geographic segmentation, Demographic, Psychographic and/or behavioral segmentation as the main approaches to segment successful their markets (Wayabila, 2020). This approach though, has been limiting customer segment development to just these considerations which does not completely bring out the entire nature of the market segments. It is also based on various assumptions that customers in the same geographical or demographic sections have the same response to given products. (Nyabwari, 2017) This is not a correct assumption. This has brought about the risk of ignoring potential markets and challenges in keeping up with the dynamic changes in the characteristics of their markets. (Yabs, 2014)

There is therefore the need to consider more data with a greater variety of business and customer attributes while carrying out market segmentation.

2.5. Linking Existing Work to My Proposed Approach

There have been various advancements in customer market segmentation. Customer demographic data from daily mobile conversation information was used to segment customers (Samuel Kamande, 2018), while data based on the average amount of goods purchased by a customer per month and the average number of customer visits per month in order to advice on a business customized marketing program for select market segments. (Dr. Manju Kaushik, 2014) However, these approaches have used limited sources of data for their market segmentation.

Most businesses in Kenya have also limited their market segmentation to only geographic, demographic, psychographic and/or behavioral segmentation. (Wayabila, 2020) This research intends to use a variety of different data features from different sources as the platform for market segmentation. This research also looks to incorporate a high dimensional clustering algorithm for the clustering process to avoid the complications that may come up with high dimensional data on ordinary clustering algorithms.

2.6. Process Model

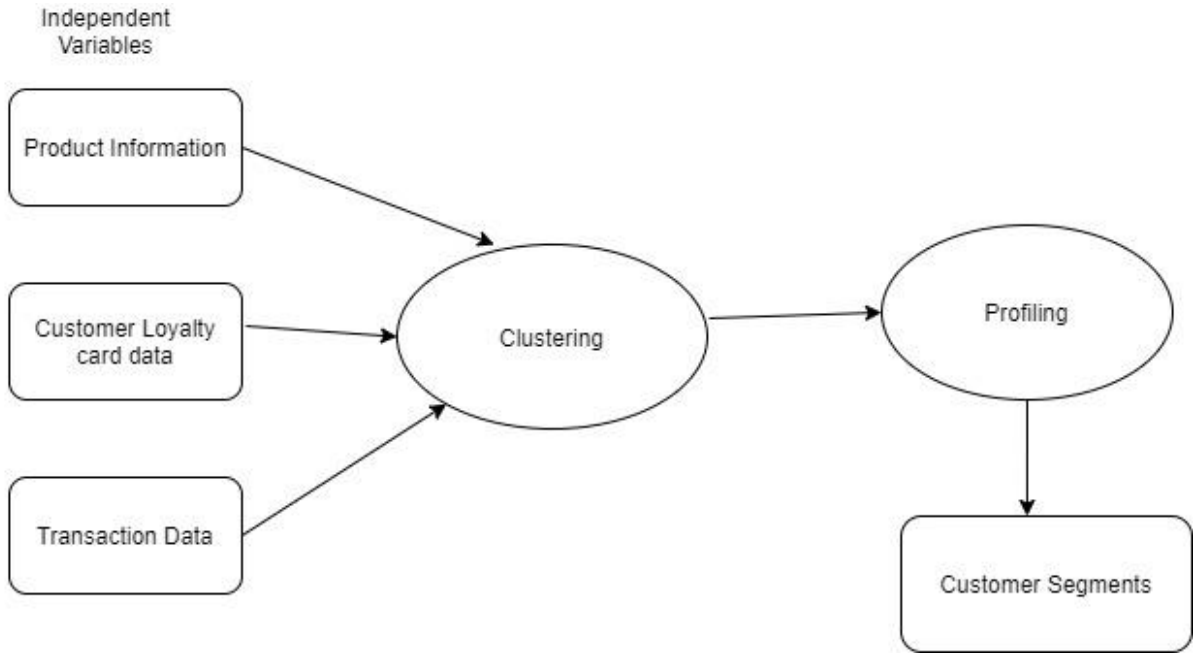

Figure 1: Process Model

CHAPTER 3: METHODOLODY

3.1. Introduction and the General Methodological Approach

The research problem to be addressed is that organizations in Kenya have been limiting their consideration to only a few factors during market segmentation. These approaches have not been completely effective, bringing about the risk of potentially ignoring various market segments and not managing to keep up with changes in customer behavior. (Yabs, 2014) This project aims to develop effective market segments through the consideration of more variated factors available in business data.

This project will have no unfair benefit to any party. The data collection methods to be used will meet validity and reliability requirements of the research. The data obtained during the research will also be used in utmost confidentiality.

3.2. Research Design

This research aims to create a machine learning model to enable businesses in Kenya to come up with up to date and effective market segments. The research design will have a mixed methods approach where I will use both qualitative and quantitative methods.

The research was divided into four major phases, these were guided by the research objectives earlier stated:

a. Data collection and analysis
b. Feature selection
c. Algorithm selection
d. Model development and Validation

3.2.1. Data Collection and Analysis

3.2.1.1. Data Collection

I adopted a qualitative approach to this phase. I looked through various papers and datasets on retail business sales data. I came across a supermarket dataset, it held data from three branches in three different cities. But the dataset only held features on the products and orders. The scope of the data was not enough for our research.

The next dataset that was assessed came from a popular retail store. It held disparate information on customer location, order information, freight information and product information. We finally settled on three datasets that gave held order data, customer loyalty

card data and data on the different products that were sold in a given business. The different attributes that were available regarding that one business and its customers made it a viable source of data for the research.

3.2.1.2. Data Analysis

Quantitative descriptive analysis was then carried out on the data collected from the pre-existing datasets. The different data was merged into one main dataset. This was in readiness for cleaning and preprocessing of the data.

Data cleaning and preprocessing was done using python pandas, NumPy and matplotlib packages to understand the overall structure of the data. This process helped in identifying patterns and correlations in the data, understanding the initial structure of the data, the null values and replicated data. We also aimed at identifying various relationships among different columns in the data in regards to sales in the business. This is because this research aims at increasing sales based on creating effective market segments based on the data in the business.

3.2.2. Feature Selection

The research employed a quantitative approach to this phase where I used a feature selection algorithm to bring out the best features to use in the dataset. These features are the ones that had the most effect on sales in the business according to the dataset. This was with an aim to prevent overfitting, and noisy data in order to ensure that the model was created with the best data from the dataset.

The data was first divided into categorical and integer values. The feature selection would be done differently for each against the sales variable. The results would then be tabulated in order to come up with the final dataset.

Integer Data

A correlation matrix was used on the integer dataset to come up with the features that had the highest correlation to Sales. The sales were as follows:

SALES 1.000000

PRICEEACH 0.657841

QUANTITYORDERED 0.551426

YEAR_ID 0.035647

MONTH_ID -0.009605

Name: SALES, dtype: float64

The PRICEEACH for an Item had the highest correlation to sales with a correlation value of 0.6578, this was followed by QUANTITYORDERED which came second with a correlation value of 0.55. The other two features had quite a low correlation compared to the first two with YEAR_ID having a correlation of 0.0356 and MONTH_ID having a negative correlation oof 0.0096.

Categorical Data

The data was first divided into the feature dataset and the target dataset. The feature dataset held all the categorical features of the cleaned dataset. These features included:

'PRODUCTCODE','CUSTOMERNAME','PHONE','ADDRESSLINE1','CITY','POSTALCODE','COUNTRY','DEALSIZE'

The dataset was then encoded using Label Encoding method for it to be fit for the feature selection algorithm.

Since our research is aimed at improving the sales of retail businesses, the SALES feature was selected as the target dataframe. It was converted into a categorical dataframe and further encoded to make it fit for the feature selection algorithm.

The research used two main feature selection algorithms. These were Chi Square algorithm and the Mutual Information Feature Selection algorithm.

The outcome of the Chi Square Algorithm on the dataset with Label Encoding was as follows:

Feature 1: 689.133762

Feature 2: 166.133044

Feature 3: 37.178916

Feature 4: 194.618424

Feature 5: 94.900497

Feature 6: 230.225534

Feature 7: 152.028044

Feature 8: 46656.412242

The outcome of the Mutual Information Algorithm was as follows:

Feature 1: 0.276338

Feature 2: 0.000000

Feature 3: 0.000000

Feature 4: 0.000000

Feature 5: 0.000000

Feature 6: 0.000000

Feature 7: 0.008426

Feature 8: 0.612279

The feature dataset was further encoded using Ordinal Encoding. The feature selection algorithms were the run again on the dataset.

The following was the results of the Chi Square algorithm on the ordinal encoded feature dataset.

Feature 1: 5661.563301

Feature 2: 48.594540

Feature 3: 8.190706

Feature 4: 22.346289

Feature 5: 28.136418

Feature 6: 41.491590

Feature 7: 10.418321

Feature 8: 546.639234

The following was the results of the Mutual Information Algorithm:

Feature 1: 0.285088

Feature 2: 0.000000

Feature 3: 0.000000

Feature 4: 0.000000

Feature 5: 0.015449

Feature 6: 0.000145

Feature 7: 0.009123

Feature 8: 0.619553

This outcome was based on the following feature information:

Feature 1: PRODUCTCODE

Feature 2: CUSTOMERNAME

Feature 3: PHONE

Feature 4: ADDRESSLINE1

Feature 5: CITY

Feature 6: POSTALCODE

Feature 7: COUNTRY

Feature 8: DEALSIZE

A separate analysis was carried out to investigate the integer-based features that most affected the sales of the company.

This was done by a correlation analysis of the integer features against the sales feature in the dataset.

The following were the results of the exercise:

SALES 1.000000

PRICEEACH 0.657841

QUANTITYORDERED 0.551426

YEAR_ID 0.035647

MONTH_ID -0.009605

3.2.3. Algorithm Selection

This phase was carried out qualitatively. The research involved qualitatively carrying out research on the machine learning algorithms that are currently used in market segmentation. Our focus was on clustering algorithms since we needed to identify patterns in our dataset without using training data.

K-Means (Chinedu Pascal Ezenkwu, 2015) and agglomerative hierarchical algorithms (Samuel Kamande, 2018) had previously been used for market segmentation. These two algorithms used the distance metric to carry out cluster formation. (Bock, 2021)

As dimensions(columns) grow in a dataset, the distance to the nearest neighbor becomes undistinguishable between the centroid and various points. This is identified as the curse of dimensionality. (M. Pavithra, 2017) This factor weakens the K-Means, Hierarchical algorithm and other distance-based clustering methods when it comes to high dimensional data.

There was therefore the need to look for an alternative option to cluster our data due to the high dimensions created in the data as the data went down the data mining pipeline.

Spectral co-clustering algorithm is an algorithm that enables the identification of clusters through similarity based on a subset of attributes in high dimensional data spaces. (Zimek, n.d.) A factor which is not affected by high dimensional data spaces. It is a form of bi-clustering algorithm which simultaneously clusters the rows and columns of a data input matrix. (Shuding Huang, 2015) The presence of column clustering would suit this research since it would establish the different clusters various customer and business features belonged to. This is unlike K-Means and Hierarchical algorithms which only cluster the rows of an input matrix. This research decided to use spectral co clustering since it could effectively cluster high dimensional data and also produced clusters by simultaneously clustering the columns of a dataset which could help us understand our data even better.

3.2.4. Model Development and Validation

3.2.4.1. Design

The following is the design that guided the model development process.

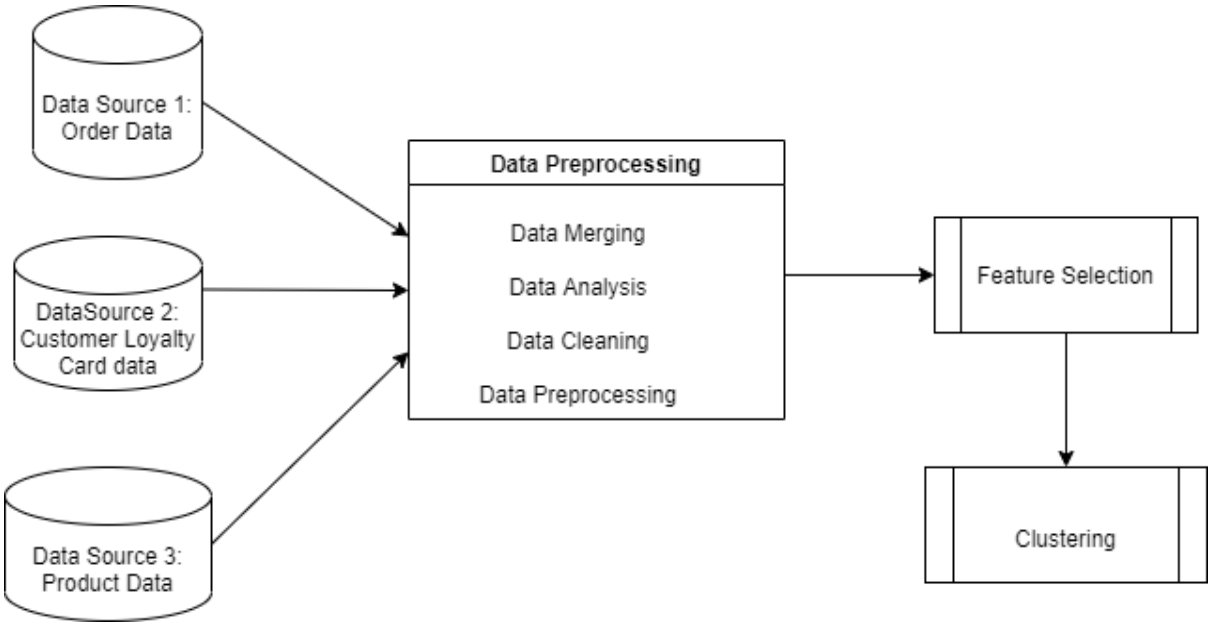

Figure 2: Model development process design

A new dataset was created from the feature set highlighted in 3.2.2 above. This would be the dataset that would be used for clustering.

The integer data in the dataset was first converted to categorical data. The categorical data was then encoded into dummy indicators. This was done using the pandas get_dummies package (Team, 2021). The data was then standardized to make it fit for clustering.

3.2.4.2. Model Design

The Spectral co-clustering algorithm discussed above was then used to cluster the data. The model of the algorithm works to simultaneously cluster the rows and columns of an input data matrix. On a bipartite graph, G=(Vr, Vc, E) where Vr is the row vertices of a dataset, Vc are the column vertices of a dataset and Eij is the edge weight between the node xi and node xj in the dataset, the algorithm looks to partition the row vertices Vr into k groups and the column vertices into l groups where k=l in order to create sub clusters. (Xiaoxiao Shi, 2010)

3.2.4.3. Validation

The Calinski_harabasz score was used to determine the appropriate clusters for the algorithm. The algorithm was also used to evaluate the model where it was used to measure the ratio

between the dispersion within the cluster and the dispersion between clusters. (learn, 2020) Better performing clusters were identified where the scores were higher. This indicated that the clusters were dense and well separated. It was a suitable option for the evaluation process since the ground truth labels in our case were are not known. (Dey, 2021)

CHAPTER 4: RESULTS AND DISCUSSION

4.1. Data Collection and Analysis

The total data collected once merged, had a total of 2823 rows and 25 different columns. From these, four columns had null values. The company had sold 109 different products across 73 different cities.

The highest sale that was done was worth 14082 which was made for 76 vintage cars done in April 2005 and the least sale made was worth 482.13 for 11 trucks and buses done in May 2005.

There were various columns that held replicable data. These columns did not have any purpose thus they were dropped. There were also columns that had sensitive data that could personally identify and individua, these columns were also removed. The columns involved: 'TERRITORY', 'STATE', 'ADDRESSLINE2', ‘PRODUCTLINE’, 'ORDERLINENUMBER', 'QTR_ID', 'MSRP', 'CONTACTLASTNAME', 'CONTACTFIRSTNAME'

Analysis of Sales Per Month

There were more sales on the months of April, May, November and October. It would be important to identify various factors in these months that caused the increase in sales compared to other months.

Most of the sales made had a value of between 1000 and 5000 with the highest being between 2000 and 4000. Most of the items bought had a value of around 100 per item.

Correlations

Price and Quantity Ordered had the highest correlations to Sales. Price had the highest correlation to the Sales outcome with a correlation of 0.657841. The quantity Ordered had the second highest correlation rate at 0.551426. The others had poor correlation with Year_Id and Month_id performing poorest with 0.035647 and -0.009605 correlation value.

4.2. Feature Selection

From the feature selection exercise explained above, we observed that the PRICE_EACH column was the feature that most affected Sales, QUANTITYORDERED feature also highly affected Sales with a correlation value of 0.55. The other two features, YEAR_ID and MONTHY_ID were not very instrumental in affecting sales.

As for the categorical data, the table below was used to summarize the results:

	Label Encoding	Ordinal Encoding
Chi Square Algorithm	Feature 1: 689.133762 Feature 2: 166.133044 Feature 3: 37.178916 Feature 4: 194.618424 Feature 5: 94.900497 Feature 6: 230.225534 Feature 7: 152.028044 Feature 8: 46656.412242	Feature 1: 5661.563301 Feature 2: 48.594540 Feature 3: 8.190706 Feature 4: 22.346289 Feature 5: 28.136418 Feature 6: 41.491590 Feature 7: 10.418321 Feature 8: 546.639234
Mutual Information Algorithm	Feature 1: 0.291104 Feature 2: 0.000000 Feature 3: 0.000000 Feature 4: 0.000000 Feature 5: 0.000000 Feature 6: 0.000000 Feature 7: 0.000000 Feature 8: 0.623955	Feature 1: 0.299570 Feature 2: 0.000000 Feature 3: 0.000000 Feature 4: 0.021951 Feature 5: 0.008777 Feature 6: 0.010569 Feature 7: 0.002295 Feature 8: 0.625836
Key	Feature 1: PRODUCTCODE Feature 2: CUSTOMERNAME Feature 3: PHONE Feature 4: ADDRESSLINE1	

	Feature 5: CITY Feature 6: POSTALCODE Feature 7: COUNTRY Feature 8: DEALSIZE	

The application of different encoding methods evidently affected the outcome of the feature selection algorithms where ordinal encoding completely changed the results of the Chi Square algorithm but only introduced a high sensitivity to the data in Mutual Information algorithm which was brought out by results in some features where none was recorded while using Label Encoding.

In the first instance where only Label Encoding was used, the Chi Square algorithm brought out the strongest features affecting sales to be: PRODUCTCODE, POSTALCODE and DEALSIZE with the DEALSIZE being the strongest with a measure of 46656.412242 and the lowest being PHONE, with 37.17.

The Mutual Information algorithm was not very sensitive to most features where it recorded zero correlation of 6 features to the Sales feature. The features that were brought out to have correlations were PRODUCTCODE and DEALSIZE. In this scenario, PRODUCTCODE had an effect on sales with a measure of 0.291104 whereas DEALSIZE had the highest measure with 0.623955.

After further encoding the feature dataset using ordinal encoding, the results of the feature selection algorithms changed. The results of the Chi Square algorithm completely changed where the features with the strongest correlations to sales became PRODUCTCODE, and DEALSIZE only. PRODUCTCODE had the highest correlation with a measure of 5661.56 and DEALSIZE was second with a measure of 546.639.

In the Mutual Information algorithm, the non-zero results were more compared to the previous Label Encoding reading where 6 features did not have a correlation. In this scenario only two features had no correlation, the other features registered correlations to the Sales feature. However, despite that change, DEALSIZE still recorded the highest correlation with a measure of 0.625836, followed by PRODUCTCODE which had a reading of 0.29957. ADDRESSLINE1 followed with a significantly low score of 0.021951. CUSTOMERNAME and PHONE recorded no correlation to the Sales feature.

At the end of the exercise, the features that most affected sales according to the algorithms were:

Chi Square (Average):

DEALSIZE: 23601.525738

PRODUCTCODE: 3175.3485

Mutual Information (Average):

DEALSIZE: 0.6248955

PRODUCTCODE: 0.295337

From the Integer data the following features had a high effect on sales:

PRICEEACH 0.657841

QUANTITYORDERED 0.551426

The highlighted features above together with the sales feature are the ones that would constitute of the dataset to be used to aid in identifying the suitable marketing segments for the business. The final dataset will therefore comprise of the following features:

- PRICEEACH
- QUANTITYORDERED
- DEALSIZE
- PRODUCTCODE

After feature selection, the conceptual framework was as follows:

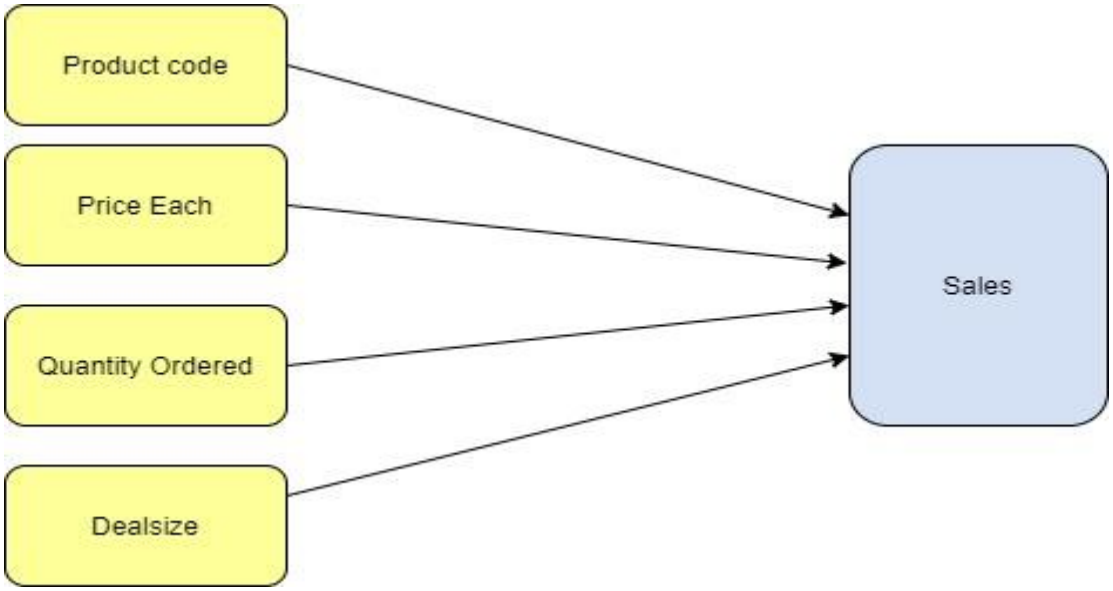

Figure 3:Conceptual Framework

4.3. Customer Segmentation Results

The aim of this phase was to identify customer clusters from the dataset. After carrying out dataset encoding, the research used the Calinski-harabasz algorithm to compare clustering scores among different cluster numbers, the best score was 3 clusters. Thus, the outcome of the clustering process was three different market segments. The research focused on the clustering of the columns of the data set since these were the features that affected sales in the business. The features were clustered as follows. This was based on the final encoded dataset where the column values had been split into dummy values for each column variable in the categorical values.

CLUSTER 1

The following were the characteristics of the first cluster:

Customers who bought products with the following codes:

CLUSTER 1
Vintage Cars:
'S18_1367', 'S18_2248', 'S18_2432','S18_2581', 'S18_2625','S18_2957','S18_3136',
'S18_3782','S18_4668', 'S18_4933', 'S50_1341'

Classic Cars:
‘S24_1046', 'S24_1444', 'S24_1628','S24_1937', 'S24_2022', 'S24_2360', 'S24_2840', 'S24_2841', 'S24_3371', 'S24_3420', 'S24_3816', 'S24_3969'
Trucks and Buses:
'S32_2206', 'S32_3207', 'S32_3522', 'S32_4289'
Trains:
'S50_1514'
Ships:
'S700_1138', 'S700_1691', 'S700_2047'
Planes:
S72_1253', 'S72_3212'
Dealsize : SMALL
Sales: 483 – 3184
Quantity Ordered: 11-30 items
Prices: 11000 - 30000 per item

Table 1:Cluster 1 characteristics

CLUSTER 2

The second cluster represented customers who held the following characteristics:

CLUSTER 2
Motorcycles
‘S10_1949', 'S10_2016', ‘S10_4698', ‘S10_4757', ‘S10_4962', ‘S24_1578', S32_1374, S32_4485, S50_4713
Classic cars
‘S12_1099', 'S12_1108', ‘S12_1666', ‘S12_2823', ‘S12_3148', ‘S12_3380', ‘S12_3891', ‘S12_4473', ‘S12_4675', ‘S24_2766', ‘S24_2887', ‘S24_3191', ‘S24_3432', ‘S24_3856', ‘S24_4048', ‘S24_4620', S700_2824

Trucks and Buses
‘S18_1097', ‘S18_1129', 'S18_1342', 'S18_1589', ‘S18_1662', ‘S18_1749', ‘S18_1889', ‘S18_1984', 'S18_2238', ‘S18_2319', ‘S18_2325', ‘S18_2795', 'S18_2870', ‘S18_2949', ‘S24_2300', ‘S18_3029', ‘S18_3140', ‘S18_3232', ‘S18_3259', ‘S18_3278', ‘S18_3320', ‘S18_3482', ‘S18_3685', ‘S18_3856', ‘S18_4027', ‘S18_4522', ‘S18_4600', ‘S18_4721', S32_1268', S50_1392
Planes
S24_1785, S24_3949, S24_4278, S700_2834, S700_3167, S700_4002
Vintage Cars
‘S24_3151', ‘S24_4258'
Ships
S24_2011, S700_1938, S700_2610,
Dealsize: Medium
Sales: 4509-14082
Quantities: 41-100 items
Prices: 41000-100000

Table 2: Cluster 2 Characteristics

CLUSTER 3

The customers in the third cluster had the following characteristics:

CLUSTER 3
Classic Cars
S12_3990, S24_2972
Vintage Cars
S18_4409,
Motorcycles
S24_2000,
Trucks and Buses

S32_2509
Planes
S700_2466,
Ships
S700_3505, S700_3962
Sales: 3185 and 4508
Quantities :31-40 items
Price: 31000 - 40000

Table 3: Cluster 3 Characteristics

From these results, the observations made are that customers in this small business can be divided using the following the following factors:

Product codes

Product quantities

Price

Amount of sales

According to (Venkatesan, 2007) the identified market segments satisfied the following factors in a bid to evaluate the effectiveness of market segments:

Measurability: The developed market segments were easy to measure and the purchasing power of the different market segment was easy to quantify.

Identifiability: Business managers could easily recognize these market segments in the marketplace.

Sustainability: The segments were large enough to ensure business profitability. They are also highly scalable ensuring business growth.

Accessibility: The identified segments could easily be identified by organization

Actionability: The customers in the different market segments would also fall in line with the core goals of the organization to increase their sales and subsequent profits.

4.4. Discussion

According to previous research done, the approaches to customer market segmentation in small Kenyan businesses has been through using customer geographical, behavioral, demographic and /or psychographic data. This approach though effective has presented challenges like running the risk of missing potential market segments since the data used did not bring out the entire nature of the market segments. (Nyabwari, 2017)

This research has identified market segments through the use of a clustering algorithm on varied features in business data. The best features in regards to sales were first identified, then the spectral co-clustering algorithm was used on these features in a bid to identify patterns in the data to create various market segments. This market segmentation process is more effective and dynamic since it has used a clustering algorithm to carry out the segmentation, and the features used for the process are the best features in the data in regards to sales in the business. The business can therefore manage to come up with better and more effective market segments. Using up to date data will also help to alleviate the risk of missing potential market segments and help keep up with changes in customer habits since the data used for market segmentation will reflect on the current market characteristics

The use of varied data features from different sources of data related to the business has enabled the research to identify other features other than geographical, demographic, psychographic and behavioral factors that could be used to identify market segments in the business. The use of the spectral co-clustering algorithm has enabled the discovery of patterns in the data that would have not been easy to spot with the naked eye, and brought automation to the market segmentation process.

This will give the business a wider view of their market and empower the business to develop more data-oriented market segments from a wider business scope compared to the use of only one or two factors.

CHAPTER 5: CONCLUSION

5.1. Achievements

The research gap identified was that the common market segmentation approaches used by small retail businesses in Kenya have not entirely brought out the nature of the customers in the business and have also led to challenges like potentially ignoring new markets and difficulty

in keeping up with changing customer needs. These approaches have also been based on the assumption that customers in the same demography or same geographic region had similar responses to a given product. An assumption that was not correct.

This research had the aim of using integrated business data to carry out market segmentation. This was in a bid to alleviate the challenges presented by the current approaches and to create more effective customer segments which stem from the consideration of more variated business data features.

The aim of the first objective was to collect disparate business data. The research collected business data on the products, customers and orders made in a small retail business. From this objective, the conclusion is that there is more data available in a retail business that can be used for market segmentation other than just geographical, behavioral, demographic and psychographic data.

The aim of the second objective was to identify the best features in regards to sales. This was done using various feature selection algorithms. The best performing features to sales in the business were identified as quantity ordered, price, deal size and product code. In conclusion, it is important to carry out feature selection since different features may have different effects on sales on a business, and identifying the best features in regards to sales performance in a business will aid in developing more effective market segments.

The aim of the last objective was to identify different clusters based on the feature set that was created. The research created three different segments each with different characteristics based on the feature set used. The market segments could then be used to identify different customers.

Previous research has based their market segmentation data on tweets about the business. It then brought out the bagging CART algorithm as the best classifying algorithm to bring out patterns in the data (Muhammad Adi Khairul Anshary*, 2016), a different research has carried out market segmentation based on daily mobile conversations in Kenya and identified the agglomerative hierarchical clustering as the best algorithm to use for the customer segmentation process (Samuel Kamande, 2018). In another scenario, data from a dataset consisting of only two attributes was used. These attributes were the average amount of goods purchased by a customer per month and the average number of customer visits per month. They then used K-Means clustering algorithm which provided high levels of accuracy in their research (Chinedu Pascal Ezenkwu, 2015).

This research seeks to use highly variated integrated data features from a business while conducting market segmentation. It also used the spectral co-clustering algorithm to identify

patterns in the data and define the consequent market segments based on the best features in the data. This is a more effective approach to market segmentation since the use of integrated data aids in bringing out patterns in the business data which one or two data factors would not have brought out, thus creating more effective market segments for the business. The use of the spectral co-clustering algorithm was good for high dimensional data which would be a challenge for other distance-based clustering algorithms. The use of up to date business data reflects on the actual business and customer characteristics and aids in keeping up with customer habit changes and prevents missing out on potential markets,.

5.2. Contributions

This study contributes to research by going into how to improve business performance and targeted sales in small retail businesses in Kenya using technology. The businesses can consider different sources of business data, integrate it, carry out feature selection and use a high dimensional clustering algorithm for market segmentation. This approach could improve the performance of small business in Kenya in terms of targeted marketing and targeted product development in response to the different identified market clusters.

The characteristics of the identified market segments in this research could then be coupled with the related business data to build identifiable markets with their own specific marketing strategies. This would then aid in reaching the different customer market segments according to their unique characteristics.

Factors that would affect the uptake of this research in small retail businesses in Kenya would be trust in the market segmentation process, ICT knowledge and skills in the business, the perceived relative advantage offered by the research to the business, the ability to run a trial version to assess business impact, management support and competitive pressure faced by the business (ODUOR, 2016).

5.3. Challenges

Obtaining customer data from small businesses in Kenya was a great challenge due to the fact that most businesses have not yet fully automated their business operations, thus a lot of transaction data is not available. As for the businesses which run automated transactions, there was the hesitation of business owners to share their transaction and customer information.

Obtaining methods of clustering evaluation where no ground truth was available was a challenge, since most evaluation algorithms use the ground truth to score the performance of a clustering algorithm.

5.4. Recommendations for Future Work

More research can be done on clustering evaluation algorithms that do not need ground truth and are good for bi clustering algorithms. There can also be more research done on how to make more data available on small businesses in Kenya for research purposes.

References

Ana maría díaz-martín, m. R. S. H. D. M. O., 2015. Using twitter to engage with customers: a data mining approach.

Arcgis pro, 2020. *How density based clustering works.* [online] available at: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-density-based-clustering-works.htm

Bock, t., 2021. *Displayr.* [online] available at: https://www.displayr.com/what-is-hierarchical-clustering/ [accessed june 2021].

Chinedu pascal ezenkwu, s. O. C. K., 2015. Application of k-means algorithm for efficient customer segmentation: a strategy for targeted customer services. *Journal of advanced research in artificial intelligence.*

Dey, d., 2021. *Geeksforgeeks.* [online] available at: https://www.geeksforgeeks.org/calinski-harabasz-index-cluster-validity-indices-set-3/

[accessed june 2021].

Dr. Manju kaushik, m. B. M., 2014. Comparative study of k-means and hierarchial clustering techniques. *International journal of software & hardware research in engineering.*

Ghose, s., n.d. *Marketing segmentation: definition, criteria and other details.* [online] available at: https://www.yourarticlelibrary.com/marketing/marketing-segmentation-definition-criteria-and-other-details/50877

Ghoshal, a., 2020. *Bagging and boosting.* [online] available at: https://www.educba.com/bagging-and-boosting/

Hall, j., 2010. *Why marketing data is important to a growing business.* [online] available at: https://www.simplybusiness.co.uk/knowledge/articles/2010/06/2010-06-03-why-marketing-data-is-important-to-a-growing-business/

Hayes, a., 2020. Business marketing essential, understanding marketing segments. *Journal of marketing.*

Learn, s., 2020. *Scikit learn.* [online] available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html [accessed june 2021].

M. F. Santos, p. C. H. Q. F. P., 2005. A clustering approach for knowledge discovery in database marketing. *Wit transactions on information and communication technologies.*

M. Pavithra, d. R., 2017. A survey on clustering high dimensional data techniques. *International journal of applied engineering research ,* 12(11).

Muhammad adi khairul anshary*, b. R. T., 2016. Tweet-based target market classification using ensemble. *J. Ict res. Appl., vol 10, no. 2,* pp. 123-139.

Muthee janet, d. K. N., 2014. Influence of entrepreneurial marketing on the growth of smes in kiambu town-cbd, kenya.

Njoroge, p. M., 2015. Marketing strategies and the performance of small and medium enterprises in matuu town, machakos.

Nyabwari, e. B., 2017. *Evaluation of the effect of market segmentation on sales performance of the banking industry;. A survey of commercial banks in kisii town , kisii county.,* s.l.: s.n.

Oduor, o. C., 2016. *The factors influencing the adoption of software as a service (saas) by small and medium size enterprises(smes): a case study of nairobi county in kenya.,* s.l.: s.n.

Optimove, 2020. *Customer segmentation via cluster analysis.* [online] available at: https://www.optimove.com/resources/learning-center/customer-segmentation-via-cluster-analysis#:~:text=in%20the%20context%20of%20customer,archetypes%e2%80%9d%20or%20%e2%80%9cpersonas%e2%80%9d.

Qi xianting, w. P., 2016. *A density-based clustering algorithm for high-dimensional data with feature selection.* S.l., s.n.

Samuel kamande, k. M. E. A., 2018. Consumer segmentation and profiling using demographic data and spending habits obtained through daily mobile conversations. *International journal of computer applications.*

Seif, g., 2018. *The 5 clustering algorithms data scientists need to know*. [online] available at: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Shamala palaniappan, a. M. C. F. M. F. R. A., 2017. Customer profiling using classification approach for bank telemarketing. *International journal on informatics vizualization,* volume vol 1.

Shuding huang, h. W. D. L. Y. Y. T. L., 2015. Spectral co-clustering ensemble. *Knowledge based systems,* volume 84, pp. 46-55.

Tarver, e., 2020. *Market segmentation.* [online] available at: https://www.investopedia.com/terms/m/marketsegmentation.asp

Team, p. D., 2021. *Pandas.* [online] available at: https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html [accessed june 2021].

Thapana boonchoo, x. A. Q. H., n.d. *An efficient density-based clustering algorithm for higher-dimensional data,* s.l.: institute of computing technology, cas, beijing 100190, china.

Venkatesan, r., 2007. *Cluster analysis for segmentation,* s.l.: university of virginia, darden business publishing.

Wayabila, m., 2020. *Effect of market segementation on customer satisfaction in kenya commercial bank, mombasa county, kenya,* s.l.: s.n.

Xiaoxiao shi, w. F. P. S. Y., 2010. *Efficient semi-supervised spectral co-clustering with constraints.* S.l., s.n.

Yabs, j. K., 2014. *Market segmentation strategies used by chloride exide kenya limited as a competitive advantage tool,* eldoret: university of eldoret - school of business and management sciences.

Zimek, p. K., n.d. Subspace clustering techniques. *Encyclopedia of database systems.*

Appendix

Sample Source Code

```
url=("../data/orders_data.csv")

order_data=pd.read_csv(url)

order_data.head()

url=("../data/product_data.csv")

product_data=pd.read_csv(url)

product_data.head()

url=("../data/product_data.csv")

product_data=pd.read_csv(url)

product_data.head()

target_df = pd.cut(x=new_df['SALES'], bins=[0, 482, 2203, 3184, 4508, 14082], labels = ['0-482','483-2203','2204-3184','3185-4508','4509-14082'])

target_df=target_df.astype('category')

label_encoder = LabelEncoder()

target_df_encoded = label_encoder.fit_transform(target_df)

target_df_encoded

from sklearn.feature_selection import SelectKBest, chi2, f_regression

chi_selection=SelectKBest(score_func=chi2,k='all')

chi_selection.fit_transform(cat_df_ordinalencoded,target_df_encoded)

from sklearn.feature_selection import SelectKBest, chi2, f_regression,mutual_info_classif

m_i_selection=SelectKBest(score_func=mutual_info_classif,k='all')

m_i_selection.fit_transform(cat_df_ordinalencoded,target_df_encoded)
```

```
m_i_selection.scores_

for i in range (2,11):

    spectral_coclustering=SpectralCoclustering(n_clusters=i,
random_state=0)

    spectral_coclustering.fit(scaled_df)

    #score=silhouette_score(data_cols_num,
spectral_coclustering.column_labels_)

    c_score=metrics.calinski_harabasz_score(data_cols_num,
spectral_coclustering.column_labels_)

    print("calinski_harabasz score for", i, "clusters is", c_score )

c1_columns = np.array(cluster_1_columns).flatten()

c1_columns = c1_columns.tolist()

c1_columns

c2_columns = np.array(cluster_2_columns).flatten()

c2_columns = c2_columns.tolist()

c2_columns
```