

CUSTOMER SEGMENTATION

A PROJECT REPORT

Submitted by

SUBHAM BANERJEE (20BAI10051)

RISHABH MATHUR (20BAI10060)

PRIYAM JAIN (20BAI10087)

YASHASWI PATEL (20BAI10327)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING WITH SPECIALIZATION IN

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING



VIT[®]
B H O P A L
www.vitbhopal.ac.in

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

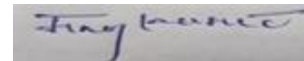
VIT BHOPAL UNIVERSITY

**KOTRIKALAN, SEHORE
MADHYA PRADESH - 466114**

APRIL 2022

BONAFIDE CERTIFICATE

Certified that this project report titled “ **CUSTOMER SEGMENTATION**”. is the bonafide work of “**SUBHAM BANERJEE (20BAI10051), RISHABH MATHUR (20BAI10060), PRIYAM JAIN (20BAI10087), and YASHASWI PATEL (20BAI10327)**” who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported at this time does not form part of any other project/research work based on which a degree or award was conferred on an earlier occasion on this or any other candidate.



PROGRAM CHAIR

Dr S Sountharajan, Associate Professor
School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY

PROJECT GUIDE

Dr. Manoj Kumar, Assistant Professor
School of Computer Science and Engineering
VIT BHOPAL UNIVERSITY

The Project Exhibition I Examination is held on **23rd April 2022**.

ACKNOWLEDGEMENT

First and foremost I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I wish to express my heartfelt gratitude to Dr S.Sountharajan, Program Chair, School of Computer Science and Engineering for much of his valuable support and encouragement in carrying out this work.

I would like to thank my internal guide Dr. Manoj Kumar, for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of Computer Science and Engineering, who extended directly or indirectly all support.

Last, but not least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

LIST OF ABBREVIATIONS

MB	Market Basket
LC	Latent Class
PAC	Programmable Automation Controller
BIC	Bayesian information criterion
AIC	Akaike Information Criterion
CSV	Comma Seperated Values
NumPy	Numerical Python
MATLAB	MATrix LABoratory
PCA	Principal Component Analysis
BLAS	Basic Linear Algebra SubPrograms
SVM	Support Vector Machine
LAPACK	Linear Algebra Package
GPU	Graphical Processing Unit
JSON	Java Script Object Notation
GUI	Graphical User Interface
SQL	Structured Query Language
OpenGL	Open Graphics Library
TkInter	TK Interface
API	Application Programming Interface
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
SciKit	SciPy Kit
SciPy	Scientific Python
LIBSVM	Library for Support Vector Machines
LIBLINER	Library for Linear Classification
AI	Artificial Intelligence

LIST OF FIGURES AND GRAPHS

FIGURE NO.	TITLE	PAGE NO.
1	Working Of Market Segmentation	14
2	System Architecture Diagram	16
3	Finding optimal Number of Clusters	17
4	Comparision of Algorithm	19
5	K - Means Clustering Perfomance Analysis	19
6	Spectral Nearest Neighbour Perfomance Analysis	20
7	Spectral RBF Perfomance Analysis	21

ABSTRACT

Customer segmentation is the division of a market into several separate groups of customers with comparable characteristics. Market segmentation is a powerful tool for defining and meeting client demands. Market Basket(MB) Analysis is performed using unsupervised machine learning techniques, the K-Means Clustering Algorithm and Hierarchical Clustering. MB Analysis is used to identify target clients who can be readily converged from a large group of customers. To enable the marketing team to devise a strategy for introducing new items to target clients who share similar interests. We have applied unsupervised learning using three different algorithms according to prediction results. We have also tried to implement an online learning model which can update it self with time to apply reinforcement learning in a clustering algorithm.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	List of Abbreviations	iii
	List of Figures and Graphs	iv
	List of Tables	v
	Abstract	vi
1	CHAPTER - 1: PROJECT DESCRIPTION AND OUTLINE 1.1 Introduction 1.1.1 Market Segment 1.1.2 Market Segmentation 1.1.3 Segmentation Variables 1.1.4 Requirements for Effective Segmentation 1.2 Motivation for the Project 1.3 Problem Statement 1.4 Objective of the work 1.4 Summary	 1 1 1 2 4 4 4 4 5
2	CHAPTER - 2: OVERVIEW OF LITERARY WORKS 2.1 Introduction 2.2 Existing Systems 2.2.1 Latent Class (LC) Cluster Analysis 2.2.2 Two Step Cluster Analysis 2.3 Proposed System 2.3.1 Working of Market Segmentation 2.4 Summary	 5 5 5 6 6 7 8 9

3	CHAPTER - 3: REQUIREMENT ARTIFACTS: 3.1 Introduction 3.2 System Configuration 3.2.1 Software Requirement 3.2.2 Hardware Requirement 3.3 Module Split Up 3.3.1 Numpy 3.3.2 Pandas 3.3.3 Matplotlib 3.3.4 Sci-Kit Learn	9 9 9 10 10 10 11 12 13
4	CHAPTER - 4: DESIGN METHODOLOGY AND ITS NOVELTY 4.1 Methodology 4.1.1 Data Collection 4.1.2. Training Module 4.1.2.1 Collecting and Analyzing of Data 4.1.2.2. Optimizing data using PAC, Vectorization and Normalization 4.1.2.3. Feature Reduction 4.1.2.4. Application of three different algorithms 4.1.2.5. Implementing a pseudo learning algorithm 4.2 Novelty 4.3 Software Architecture Diagram Fig 3. Finding Optimal Number of Clusters	13 13 13 13 14 15 16 17
5	CHAPTER - 5: TECHNICAL IMPLEMENTATION & ANALYSIS 5.1 Outline 5.2 Performance Analysis (Graphs/Charts)	17 17 18

	Fig 1. Comparison of Algorithm Fig 2. K - Means Clustering Fig 3. Spectral Nearest Neighbour Fig 5: Spectral RBF	19 20 21
6	CHAPTER - 6: PROJECT OUTCOME AND APPLICABILITY 6.1 Outline 6.2 Key Implementations Outlines of the System 6.3 Project Applicability on Real-World Applications	22 22 22 22
7	CHAPTER-7: CONCLUSIONS AND RECOMMENDATION 7.1 Outline 7.2 Limitation/Constraints of the System 7.3 Future Enhancements 7.4 Inference	23 23 23 24 24
	References	24

1.PROJECT DESCRIPTION AND OUTLINE

1.1. INTRODUCTION

Customer - relationship management and maintenance have always been critical in providing business knowledge to firms in order to create, manage, and grow beneficial long-term customer connections. In today's world, recognising consumers as an organisation's most valuable commodity is becoming especially prominent.

Customer segmentation is the process of categorising a company's customers into different groups based on their shared characteristics. The purpose of customer segmentation is to determine how and where to interact with the customer base in each category in order to optimise each customer's value to the company. Marketers may interact with each client in the most effective way with accurate customer segmentation. Customer segmentation aids in the identification of customers with varying tastes, expectations, wants, and traits.

1.1.1 Market Segment

A market segment is a collection of people, businesses, or organisations with similar interests, features, or characteristics. It's possible that the customer groupings have comparable requirements, interests, and expectations. Businesses must figure out how to separate and differentiate their segments in the most effective way possible. Once the segments have been determined, they must tailor their offers to meet the needs of each one.

1.1.2 Market Segmentation

The process of finding market segments and splitting a large client base into sub-groups of present and prospective customers is known as market segmentation. Market segmentation is a consumer-focused technique that may be used to virtually any market. Researchers often seek for shared traits such as same demands, common hobbies, similar lifestyles, or even similar demographic profiles when separating or segmenting audiences. So, market segmentation assumes that different segments require different marketing programmes, as diverse customers are usually targeted through different offers, prices, promotions, distributions or some combination of marketing variables

Marketing managers may gain a better grasp of their clients' requirements and wants by segmenting the market. This allows them to more precisely and ethically customise the company's marketing operations to the preferences of individual clients. Segmentation marketing helps companies meet and surpass their consumers' expectations. It may also enable them to assess the strengths and weaknesses of their competition. They may identify business prospects in markets that were underserved in this way. Customer segmentation helps marketers to take a more methodical approach to future planning. As a result, marketing resources are better used, leading in the construction of a more finely tuned marketing campaign.

1.1.3 Segmentation Variables

The conventional factors that may be utilised for market segmentation can be classified into four categories:

- Demographic
- Geographic
- Psychographic
- Behavioural

- A. **DEMOGRAPHIC SEGMENTATION** : Demographic segmentation is the process of breaking a market into recognisable categories based on physical and factual data. Age, gender, income, employment, marital status, family size, race, religion, and country are some of the demographic characteristics to consider. Because demographic characteristics are generally straightforward to assess, these segmentation approaches are a common means of segmenting client markets.
- B. **GEOGRAPHIC SEGMENTATION** : Geographic segmentation is the process of identifying prospective markets based on their location. Climate, terrain, natural resources, and population density, among other geographic characteristics, may be considered in this segmentation strategy. Because one or more of these characteristics can distinguish clients from one region to the next, markets can be separated into regions.
- C. **PSYCHOGRAPHIC SEGMENTATION** : Psychographic segmentation may be used to divide markets into groups based on personality traits, values, motivations, interests, and lifestyles. A psychographic dimension can be used by itself or in combination with other segmentation factors to segment a market. When purchasing habits are linked to a customer's

personality or lifestyle, psychographic variables are employed. Different customers may react differently to a company's marketing activities.

- D. **BEHAVIOURAL SEGMENTATION** : The term "behavioural segmentation" refers to the division of a market into segments based on individual purchasing habits. The advantages desired from the product, as well as the identification of certain buying behaviours, such as shopping frequency and volume of purchase, are all evident in behaviour-based segmentation.

1.1.4. Requirements for Effective Segmentation

A market can be split in a number of different ways. Not all market segmentations, however, are effective. Market segments must be meaningful, and they must be relevant to the product being promoted. The following features must be present in the market segments:

- Measurability
- Substantiality
- Accessibility
- Actionability
- Differentiability

- A. **MEASURABILITY** : The size and purchasing power of the sector must be quantifiable. It must be feasible to acquire actual data about the market's varied features. Travellers who earn \$100,000 per year, for example, account for 42% of all passengers. Businesses' marketing plans and methods would be more effective if they had reliable data about their target groups.
- B. **SUBSTANTIALITY** : The brand that would want to penetrate the market that is Substantial in number. Customer profiles should clearly be defined by gathering data of their age, gender, job, socio-economic status and purchasing power.
- C. **ACCESSIBILITY** : The customers and consumers should easily be able to reach at an affordable cost. This can help in making ads more profitable by determining different target markets for particular markets.
- D. **ACTIONABILITY** : This refers to the extent to which successful programmes may be altered to appeal to and serve relevant audiences. A small airline, for example, may be able to identify multiple market niches, but its personnel and financial resources may restrict its capacity to implement diverse marketing programmes effectively.

- E. DIFFERENTIABILITY : When segmenting a market, it should be made sure that different target markets respond differently to different marketing strategies. If a business is only targeting one market segment, then this might not be an issue.

1.2 MOTIVATION FOR THE PROJECT

While mass marketing techniques can still provide results, assuming that everyone would want to buy what is being sold is a time-consuming, inefficient, and costly approach. Market segmentation may aid in the identification and comprehension of target audiences and ideal consumers. This allows you to pinpoint the correct market for your items and then more effectively focus your marketing. Market segmentation may also assist businesses in developing goods that better fulfil their consumers' demands. Products can be developed to appeal to the demands of the major market segment, and other products customised to different segments of the client base can be developed.

1.3. PROBLEM STATEMENT

Businesses may not be able to satisfy all of its consumers on a consistent basis. It may be challenging to suit each customer's specific expectations. Because no two people have the same preferences, a single product seldom meets everyone's needs. As a result, many businesses will often use an approach known as target marketing. This approach entails segmenting the market and providing products or services for each segment.

1.4. OBJECTIVE OF THE WORK

- I. The aim of the project is to create an online learning model which can update itself and learn with time.
- II. Application should improve itself with time the help of Machine Learning and can increase its efficiency and accuracy.
- III. To implement a real time Customer Segmentation model that updates after 50 people have been registered on the basis of their age, salary and spending score.
- IV. To design the model that can correctly group different people in a cluster according to their age, salary and spending score and predict how much they can they usually spend and what all the target segment usually buy.

1.5 SUMMARY

Businesses may not be able to satisfy all of its consumers on a consistent basis. It may be challenging to suit each customer's specific expectations. Because no two people have the same preferences, a single product rarely meets everyone's needs. Customer segmentation aids in the identification of clients who have different tastes, expectations, demands, and characteristics. A market segment is a group of people, firms, or organisations who have similar interests, qualities, or behaviours. Market segmentation is the process of identifying market segments and dividing a huge consumer base into sub-groups of current and prospective customers. Market segmentation is a customer-focused strategy that may be used to almost any market. The traditional factors that can be used for market segmentation are divided into four groups that are Demographic, Geographic, Psychographic and Behavioural. Five requirements for effective customer segmentation is Measurability, Substantiality, Accessibility, Actionability, Differentiability

2. OVERVIEW OF LITERARY WORKS

2.1. INTRODUCTION

The domain study that we performed for the project was primarily concerned with comprehending and understanding Clustering and Dimensionality Reduction

2.2. EXISTING SYSTEM

We went through additional similar attempts that are implemented in the realm of Customer Segmentation Using Online Approach in the Literature survey. The following are descriptions of each of the project works.

2.2.1. Latent Class (LC) Cluster Analysis

Analysts can choose any number of segmentation inputs or indicators as well as variables (such as demographics) for the model using LC cluster analysis, as implemented by Latent GOLD® 4.5 (Statistical Innovations Inc., 2008). In an LC cluster model, the indicators are dependent variables that are used to define or assess the latent classes. They are the main factors that influence segmentation. Covariates, which might be demographics or crucial outcome factors like purchase intent for a new product, are secondary drivers. In the study, covariates might be handled as active (allowing them to impact clustering) or inactive (serving merely as profiling variables).

The Bayesian information criterion (BIC) is one of the methods available in LC cluster analysis for selecting cluster models. The BIC is a statistical measure of how well a model describes data. The lower the BIC, the better.

LC cluster analysis offers the most convincing methodological benefit since, unlike the other segmentation approaches presented in this study, it is based on probability modelling. As a result, one can argue that these segments are "actual" rather than merely an intriguing way of looking at the data.

However, LC cluster analysis takes longer to perform than other methods, especially when dealing with large data sets with thousands of respondents. Using a high-speed computer, scientists have experienced run times of many hours for huge, difficult segmentation tasks. To assist the analyst sift through the variety of alternatives available, LC cluster analysis necessitates strong statistical understanding. Because LC cluster analysis can handle so many variables, it's easy to go overboard with segmentation inputs. Excessive intricacy makes it more difficult to comprehend the segmentation solution.

2.2.2. Two Step Cluster Analysis

Factor scores or individual attributes can serve as input into TwoStep cluster analysis. Additionally, TwoStep can handle categorical variables, such as demographics (e.g., gender,

ethnicity) rated on a satisfaction scale. For the current analysis, the 33 individual attributes, classified as categorical, were used as the segmentation variables.

The analyst can either provide the number of clusters or have the algorithm choose the number based on the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC). There's also a way to deal with responders who don't fulfil the requirements for any of the clusters. These "outlier" responders are put together to prevent them from being profiled further.

In comparison, two-step cluster analysis provides several advantages. One advantage is the variety of cluster sizes available. Factor segmentation and k-means are known for producing clusters of comparable size. TwoStep produces clusters with a wider size range. Having a segmentation solution with different-sized clusters has greater face validity. Categorical qualities, on the other hand, may be given as such in TwoStep. This can help to improve segment separation and make the findings easier to comprehend.

The TwoStep approach, however, has several drawbacks. TwoStep is influenced by the order of the records in the data set, just like k-means clustering. Sorting the data records in several ways can help the analyst understand how the cluster profiles change with different orderings. In addition, respondents with any missing data are eliminated from the study completely. If a substantial percentage of respondents ignore or refuse to answer essential segmentation questions, this might reduce the sample size available for segmentation.

2.3. PROPOSED SYSTEM

The suggested system is an unsupervised learning based Customer Segmentation Model with an Online Approach that recognizes and segments different customer according to their buying and spending habits, age, income etc. Using an Online approach, the model updates and learns all by itself with time and makes clusters of different market segments with utmost accuracy. Unsupervised learning has been applied using three different algorithms to compare and implement

the best algorithm according to the prediction results. The goal of segmenting customers is to decide how to relate to customers in a business.

We'll utilize unsupervised learning to categorize the clients into clusters based on individual characteristics like age, gender, area, and hobbies. The following steps will be used to apply K-means clustering or hierarchical clustering:

- a. Business Care
- b. Data Preparation
- c. Segmentation with K-Means Clustering
- d. Tuning the Hyperparameters

2.3.1 Working of Market Segmentation

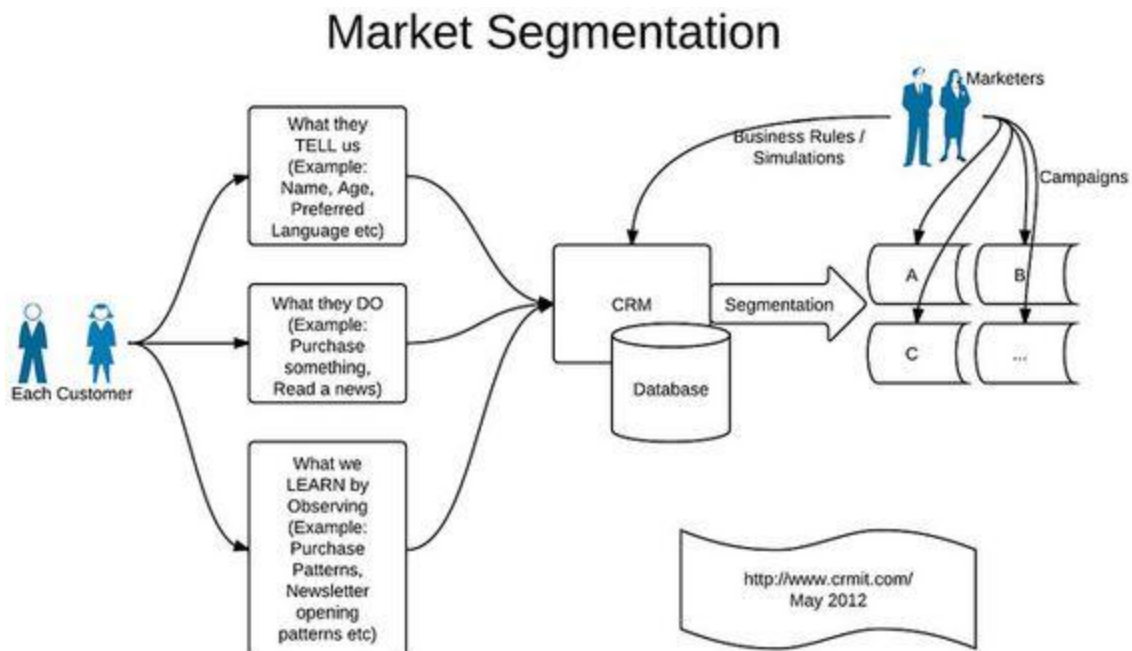


Fig 1: Working Of Market Segmentation

2.4. SUMMARY

Because of the intense rivalry in the business sector, businesses have had to improve their profitability and business throughout time by meeting client requests and attracting new customers based on their wants. Client identification and meeting individual customer wants is a difficult process. This is due to the fact that clients differ in terms of their wants, desires, preferences, and so on. Customer segmentation, as opposed to a "one-size-fits-all" strategy, separates consumers into groups with similar features or behavioural traits.

3. REQUIREMENT ARTIFACTS

3.1. INTRODUCTION

The subject investigation that we implemented for the research was essentially concerned with fully comprehending neural network models.

3.2. SYSTEM CONFIGURATION

3.2.1. Software Requirements:-

OPERATING SYSTEM: Windows, Mac, Linux

LIBRARIES:

- Numpy
- Pandas
- sklearn.cluster (import K-means)
- matplotlib
- Sklearn.preprocessing (import Standard Scaler, Normalize)
- sklearn.decomposition (import PCA)
- sklearn.clustering (import spectral clustering)

3.2.2. Hardware Requirements:-

RAM: Minimum 8GB or higher

GPU: 4GB dedicated

PROCESSOR: Intel Pentium 5 or higher

HDD: 10GB or higher

MONITOR: 15" or 17" colour monitor

MOUSE: Scroll or Optical Mouse or Touchpad

KEYBOARD: Standard 110 keys keyboard

3.3. MODULE SPLIT UP

3.3.1. Numpy (Numerical Python)

NumPy is a Python library that adds support for massive, multidimensional arrays and matrices, as well as a vast number of high-level mathematical functions to operate on these arrays.

FEATURES: NumPy is designed to work with Python's CPython reference implementation, which is a non-optimizing bytecode interpreter. Algorithms written for this version of Python are frequently much slower than compiled equivalents. NumPy tackles the slowness issue in part by offering multidimensional arrays as well as functions and operators that operate effectively on arrays, which necessitates rewriting some code, primarily inner loops, in NumPy. Because they are both interpreted, NumPy in Python provides functionality comparable to MATLAB, and they both allow the user to construct fast programs as long as most operations work on arrays or matrices rather than scalars. In comparison, MATLAB has a plethora of extra toolboxes, most notably Simulink, but NumPy is inextricably linked with Python, a more current and comprehensive programming language. Additionally, there are additional Python programmes available; SciPy is a library that adds more MATLAB-like capability, and Matplotlib is a plotting tool that gives MATLAB-like plotting functionality.

Both MATLAB and NumPy internally rely on BLAS (Basic Linear Algebra SubPrograms) and LAPACK (Linear Algebra Package) for efficient linear algebra processing.

LIMITATIONS: Implanting or adding entries to an array is more difficult than it is with Python's lists. To expand arrays, the `np.pad(...)` procedure produces new arrays with the specified form and padding values, copies the given array into the new one, and returns it. The `np.concatenate([a1,a2])` action in NumPy does not truly link the two arrays, but instead returns a new one that contains the elements from both given arrays in order. Reshaping the dimensionality of an array using `np.reshape(...)` is only possible if the array's element count does not change. The fact that NumPy arrays must be views on contiguous memory buffers causes these conditions.

3.3.2. Pandas

Pandas is a data manipulation and analysis software package for the Python programming language. It includes data structures and methods for manipulating numerical tables and time series, in particular. Pandas is mostly used for data analysis and related tabular data manipulation in Dataframes. Pandas supports importing data from comma-separated values (CSV), JSON, Parquet, SQL database tables or queries, and Microsoft Excel. Pandas supports a wide range of data manipulation operations, including merging, reshaping, and selecting, as well as data cleaning and wrangling. Many similar aspects of working with Dataframes that were established in the R programming language were introduced into Python with the development of pandas. The pandas library is based on NumPy, a Python library geared toward effectively working with arrays rather than the characteristics of working with Dataframes.

FEATURES:

- DataFrame object with integrated indexing for data manipulation.
- Reading and writing data between in-memory data structures and several file formats with tools.
- Data alignment and handling of missing data in a unified manner.
- Data sets can be reshaped and pivoted.

- Slicing of big data sets based on labels, clever indexing, and subsetting
- Inserting and deleting columns in a data structure.
- Split-apply-combine procedures on data sets can be performed by grouping by engine.
- Merging and combining data sets.
- To work with high-dimensional data in a lower-dimensional data structure, use hierarchical axis indexing.
- Date range generation and frequency conversions, moving window statistics, moving window linear regressions, date shifting and lagging are all examples of time series capability.
- Filtering of data is available.
- Critical code paths are written in Cython or C, and the library is heavily optimised for efficiency.

3.3.3. Matplotlib

Matplotlib is a graphing package for Python with NumPy, the Python numerical mathematics extension. It provides an object-oriented API for embedding charts into applications utilising GUI toolkits such as Tkinter, wxPython, Qt, or GTK. There's also a procedural "pylab" interface built on a state machine (like OpenGL) that's meant to look like MATLAB, however it's not recommended

TOOLKITS : Matplotlib's functionality may be extended with a variety of toolkits. Some are standalone downloads, while others come packaged with the Matplotlib source code but rely on third-party libraries.

- Basemap: map projections, coastlines, and political boundaries are all plotted on the basemap.
- Cartopy is a mapping library that supports arbitrary point, line, polygon, and picture transformations as well as object-oriented map projection definitions. (Matplotlib version 1.2 and up)
- Excel tools are programmes that allow you to exchange data with Microsoft Excel.
- GTK tools: a user interface for the GTK library Qt

- 3-D graphs using Mplot3d
- Natgrid: a natgrid library interface for gridding unevenly spaced data.
- export to Pgfplots for easy inclusion into LaTeX documents using tikzplotlib (formerly known as matplotlib2tikz)
- Seaborn is a Matplotlib-based API that provides sensible plot style and colour defaults, defines simple high-level methods for popular statistical plot kinds, and interacts with Pandas' functionality.

3.3.4. SciKit-Learn

SciKit-Learn, also known as sklearn, is a free Python machine learning package. Support-vector machines (SVM), random forests, gradient boosting, k-means, and DBSCAN are among the classification, regression, and clustering algorithms included, and it is meant to work with the Python numerical and scientific libraries NumPy and SciPy.

Scikit-learn is mostly built in Python, and it heavily relies on NumPy for high-speed linear algebra and array operations. In addition, to boost performance, some key algorithms are written in Cython. A Cython wrapper around LIBSVM implements support vector machines; a similar wrapper around LIBLINEAR implements logistic regression and linear support vector machines. It may not be possible to expand these methods with Python in such instances. Many other Python libraries, such as Matplotlib and plotly for graphing, NumPy for array vectorization, Pandas dataframes, SciPy, and others, work well with Scikit-learn.

4. DESIGN METHODOLOGY AND ITS NOVELTY

4.1. METHODOLOGY

4.1.1. Data Collection

The first step is the data-set collection. We started the process of gathering and measuring information on our variables of interest by collecting the dataset from kaggle.

For the real world application of the model we have planned to collaborate with small businesses and startups to know about their market base and increase the efficiency of the product.

4.1.2 Training module

We have trained the model by applying UNSUPERVISED MACHINE LEARNING using four different algorithms to compare and implement the best algorithm according to the prediction results.

UNSUPERVISED MACHINE LEARNING : The use of artificial intelligence (AI) systems to find patterns in data sets including data points that are neither categorised nor labelled is known as unsupervised learning. Unsupervised learning allows the system to recognise patterns in data sets without the assistance of a human. Even if no categories are specified, an AI system will categorise unsorted data according to similarities and differences in unsupervised learning. Compared to supervised learning systems, unsupervised learning algorithms can handle more complex processing tasks. Additionally, one method of putting AI to the test is to put it through unsupervised learning.

To develop such the model, the following stages must be completed:

1. Collecting and analysing data
2. Optimising data using PAC, Vectorization and Normalisation
3. Feature Reduction
4. Application of Three different algorithms
5. Implementing a pseudo learning algorithm

4.1.2.1. COLLECTING AND ANALYSING OF DATA

The First step is to collect the data from Kaggle as collecting it manually or creating our own dataset might result in missing information and errors. It is then analysed and manipulated using pandas. The data is then processed and optimised according to our training model

4.1.2.2 OPTIMISING DATA USING PAC, VECTORIZATION AND NORMALISATION

This step is done to optimize the code to function efficiently with large amounts of data by using several cores available on current CPUs to parallelize our programmes. Fundamentally, this continues to improve the program's runtime and memory allocations.

4.1.2.3. FEATURE REDUCTION

PCA was used for feature reduction as it reduced the number of dimensions, making the data less sparse and statistically significant. It cut down on the amount of time and storage space needed. Reduced the amount of time spent training. Assisted with avoiding the dimensionality curse. Reduced overfitting which improved generalisation.

4.1.2.4. APPLICATION OF THREE DIFFERENT ALGORITHM:

Three alternative algorithms were used in total, and the one that produced the best results was picked. The three techniques that were employed were :-

- K-means
- Closest neighbour spectral clustering
- RBF spectral clustering

K - Means was the algorithm that was chosen as it was easier to implement and gave us the best result.

4.1.2.5. IMPLEMENTING A PSEUDO LEARNING ALGORITHM

Our own model was tested using an online clustering technique from K-Means that can learn in the real world to improve clusters.

It gathered the data collected by the application and updated the algorithm for each group of 50 records received by our updating algorithm. Due to a dearth of study in this area, data was retrained rather than updated using the datasets that were already available. This is why it felt more like we were creating a fictitious online algorithm.

4.2. NOVELTY

- This project will help small business and startups cut costs in marketing.
- It will eliminate the need of human labour.
- This Model will help by highlighting the market a particular product caters to

- Will provide easy and cheap solution
- Increase the efficiency of the business

4.3. SYSTEM ARCHITECTURE DIAGRAM

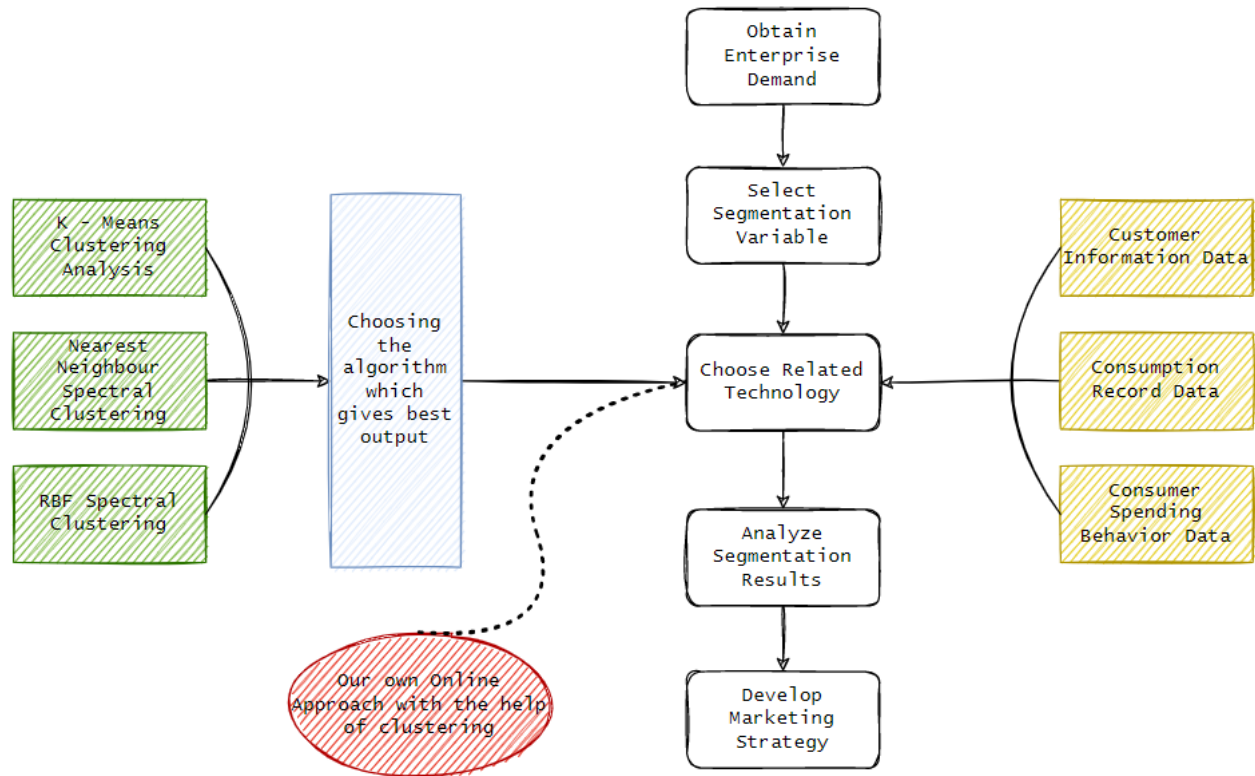


Fig 2 : System Architecture Diagram

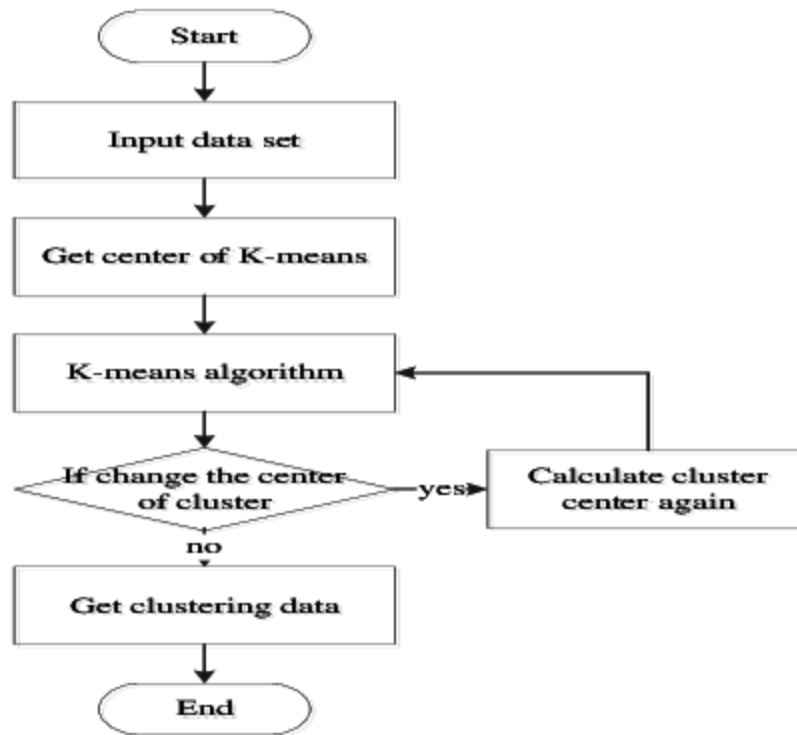


Fig 3: Finding Optimal Number Of Clusters

For K means clustering, an optimal K value, ie. optimal number of clusters were need. We found it by changing the centroids everytime as shown n the graph as shown in the flowchart above.

Chapter 5.

TECHNICAL IMPLEMENTATION & ANALYSIS

5.1. Outline

Our Customer Segmentation model with our own online approach was designed with the help of Mall Dataset and thus has been designed for Mall Customer Segmentation. Our results demonstrate the importance of Customer Segmentation for a business. Customer segmentation is critical for optimising marketing strategies, maximising a customer's worth to your company, and improving customer experience and happiness. With the world's infinite diversity of individuals and

characteristics, a potential market is rarely singled out or simply defined. To ensure that messages are both effective (attractive, action-promoting) and appropriate, it is necessary to understand the desired target consumer base (non-offensive, timely, and relevant).

The segmentation process starts with grouping customers and potential customers into customer segments with similar characteristics so that you can communicate with all of the people in that segment quickly, effectively, and with a sense of personal attention without having to contact each person individually. As a result, your marketing strategies will become more successful and efficient, saving you time and money while also increasing the rewards.

Three different algorithms were implemented for the purpose of finding the best one among them. The algorithms that were implemented were, Kmeans Clustering, Closest neighbour spectral clustering, RBF spectral clustering.

K- MEANS CLUSTERING : When you have unlabeled data, K-means clustering is a sort of unsupervised learning (i.e., data without defined categories or groups). The purpose of this technique is to locate groups in the data, with K representing the number of groups.

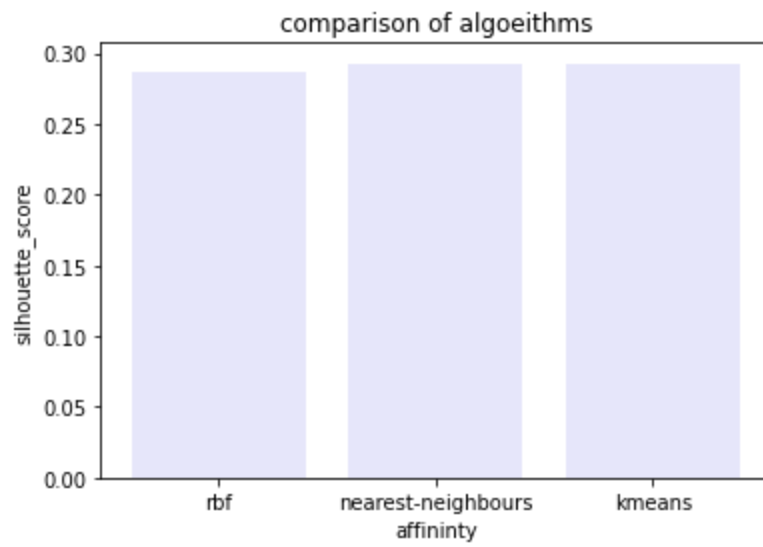
CLOSEST NEIGHBOUR SPCTRAL CLUSTERING : Spectral clustering is a technique that has its roots in graph theory and is used to discover communities of nodes in a network based on the edges that connect them. The approach is adaptable, allowing us to cluster data that isn't graphed.

RBF SPECTRAL CLUSTERING: RBF Clustering is simply Spectral Clustering with RBF Kernel. The RBF kernel, or radial basis function kernel, is a widely used kernel function in many kernelized learning techniques. It's especially popular in support vector machine classification.

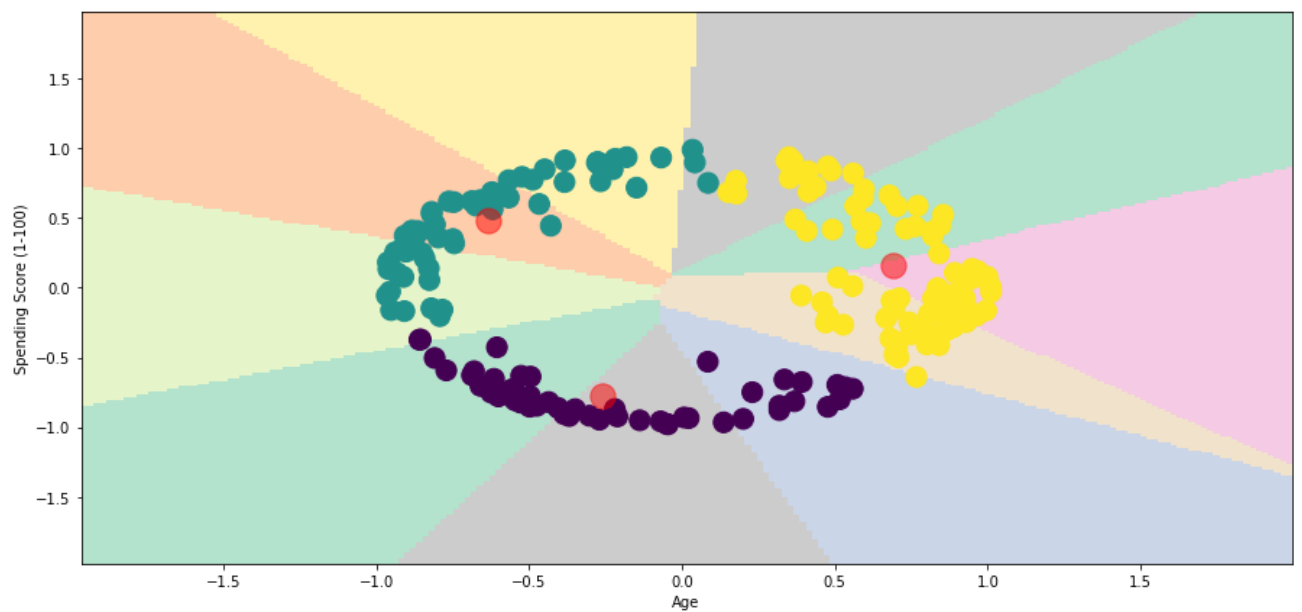
According to each algorithms Performance Analysis, the best among them is chosen. In this case its K- Means Clustering which is easier to implement and gave the desired results efficiently.

5.6. Performance Analysis (Graphs / Charts)

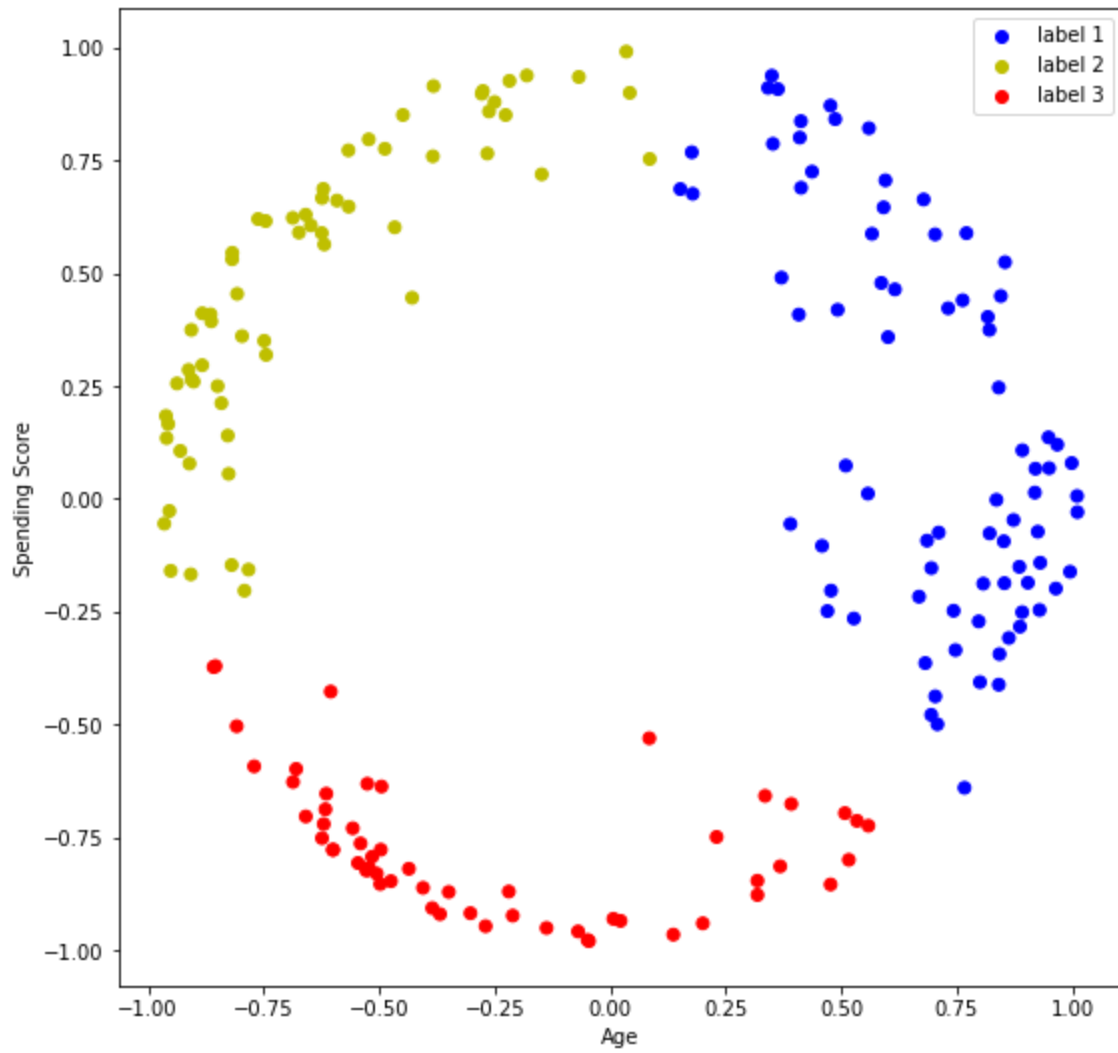
A. Fig 1: Comparison of Algorithms



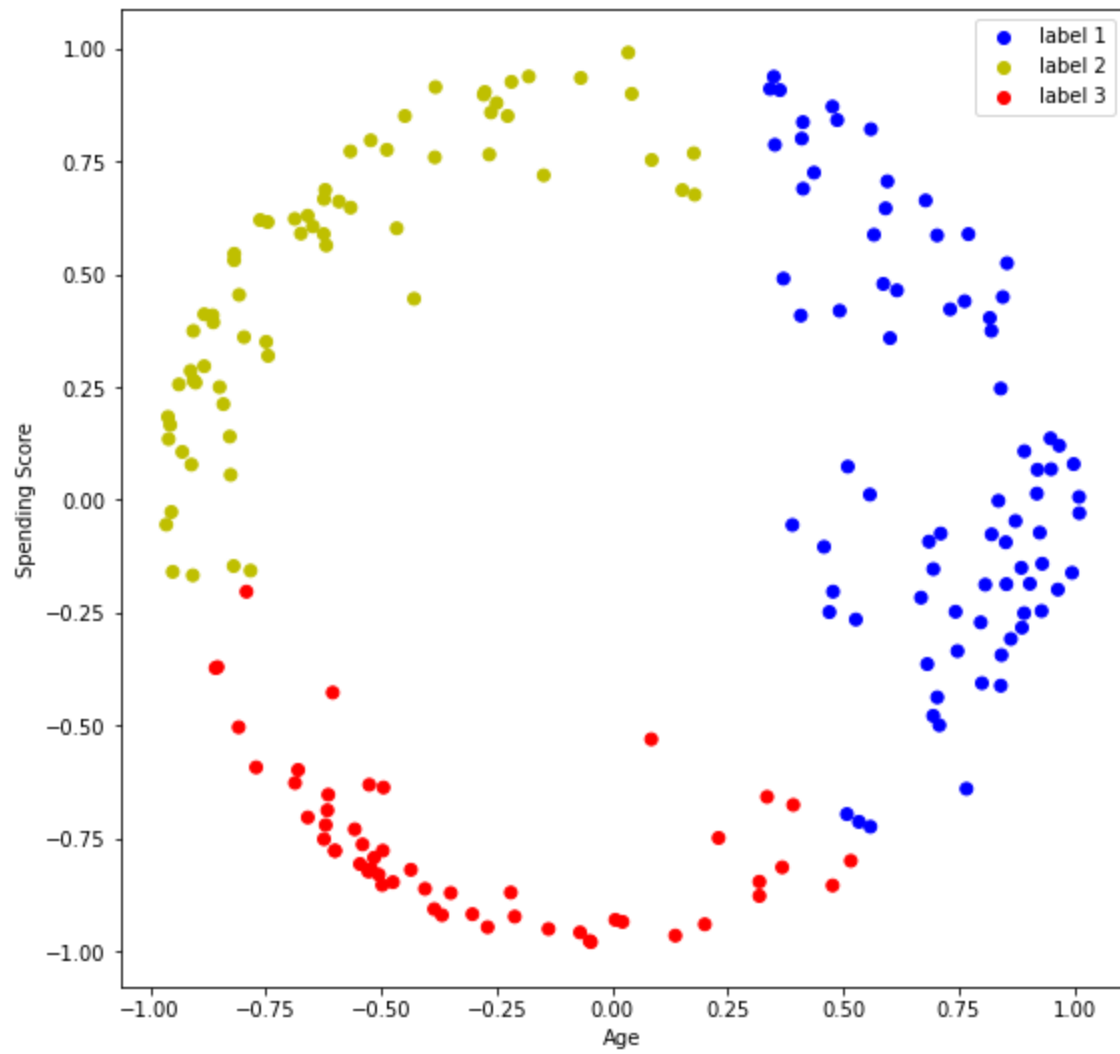
B. Fig2: K-Means Clustering



C. Fig 3: Spectral Nearest Neighbour



D. Fig 4: Spectral RBF



CHAPTER - 6:

PROJECT OUTCOME AND APPLICABILITY

6.1 Outline

The process of grouping customers into sections of individuals who share common characteristics is called Customer Segmentation. This segmentation enables marketers to create targeted marketing messages for a specific group of customers which increases the chances of the person buying a product. It allows them to create and use specific communication channels to communicate with different segments to attract them. A simple example would be that the companies try to attract the younger generation through social media posts and older generation with maybe radio advertising. This helps the companies in establishing better customer relationships and their overall performance as an organisation.

6.2. Key Implementations Outlines of the System

a. Implementation of best algorithm

b. Formation of clusters

6.3 Project Applicability on Real-World Applications

As of now the model we have created is designed specially for Shopping-Malls and caters to its need of specific customer segmentation therefore providing them with a accurate information of the demography of their customers.

But we can alter the inputs of the model making it useful to a wide variety of business and startups in knowing how suitable theirs product is in the market and which market segment they should specially target . In short , we can increase the efficiency of a product and a company .

Segmentation breaks the market down by categorizing consumers as impulse shoppers, local customers, bargain hunters, wealthy seniors and so on. Therefore we can cut costs in marketing as we can specific advertisements targeting the right market segment, with ads tailored to suit people in that segment . This means shaping your marketing based on the kind of company you're dealing with, such as a non-profit, a for-profit corporation or an independent contractor.

CHAPTER-7:

CONCLUSIONS AND RECOMMENDATION

7.1. Outline

The initial setup of customer segmentation for businesses can be a hurdle. Identifying the need for customer segmentation is the first step towards implementing a process that aligns with your overall business plan. When businesses don't have an effective customer segmentation process, they might find themselves providing the same service level for all customers and all products without focusing on the top-level customers or products that bring in the best margins. In order to maximize productivity and profitability, customer segmentation helps businesses apply the 80/20 rule, instead of spreading themselves thin by trying to provide the same service level for every customer, regardless of whether they are a top customer or not.

7.2. Limitation/Constraints of the System

One of the biggest issues with customer segmentation is data quality. Inaccurate data in source systems will usually result in poor grouping. For example, customers who are individuals, attributes like age, gender, and marital status are frequently used. If these attributes are not maintained properly, the segments will be inaccurate and as a result, the information will likely be less useful. If the users do not feel comfortable with the quality of the data, they are likely not going to use the segments. Data quality issues also arise from a lack of maintenance and regular cleansing to ensure accuracy.

7.3. Future Enhancements

The growth of the schooling photo length and accuracy betters itself at some stage in schooling and actual time implementation (because the variety of parameters to gain knowledge increases). However, schooling snap shots with massive length is without delay proportional to the computational energy of the system.

7.4. Inference

Due to the limited computational power of our laptop we were able to compute the result with accuracy of 89% but a person with a better performing engine can yield better results.

REFERENCES

1. D. P. Yash Kushwaha, "Customer Segmentation using K-Means Algorithm," 8th Semester Student of B.tech in Computer Science and Engineering
2. E. A. Onur DOGANI, "CUSTOMER SEGMENTATION BY USING RFM". [3] C. M. S. R. a. K. V. N. T. Sajana, "A Survey on," in Indian Journal of Science and Technology, Volume 9, Issue 3,, Jan 2016
3. A. B. P. E. Shreya Tripathi, "Approaches to," in International Journal of Engineering and Technology, Volume 7, 2018.
4. https://www.researchgate.net/figure/Basic-process-of-customer-segmentation_fig1_329372886
5. <https://www.yieldify.com/blog/types-of-market-segmentation/>

6. https://www.researchgate.net/publication/319085560_Market_Segmentation_Targeting_and_Positioning
7. https://www.researchgate.net/publication/230557972_Approaches_to_Customer_Segmentation

Book Name:

1. Advances in K-means Clustering: A Data Mining Thinking, Book by Junjie Wu
2. Data Clustering: Theory, Algorithms, and Applications, Book by Chaoqun Ma, Guojun Gan, and Jianhong Wu
3. Spectral Clustering and Biclustering: Learning Large Graphs and Contingency Tables, Book by Marianna Bolla
4. Market Segmentation: How to Do It and How to Profit from It, Book by Malcolm McDonald