A PROJECT REPORT

on

Customer Segmentation using Machine Learning

Submitted by

Group Id:55901

Chauhan Vimarsh A.(160840107007)

Patel Darshankumar R. (160840107031)

Patel Monilkumar D. (160840107037)

Rana Jecky A. (160840107051)

In partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

Computer Science and Engineering

R. N. G. Patel Institute of Technology,

Isroli - Bardoli

Gujarat Technological University,

Ahmedabad

November 2019

ABSTRACT

In recent years, enterprises apply many strategies in order to have a better customer relationship. Identifying the potential customer is the main factor that affect the enterprises.Since every customer has different preferences it is necessary to mine the customer data to gain meaningful insight.Thus it is necessary to group the customer based on their buying behavior, geographic, product-related approach,etc.

In our proposed system, we will segment the potential customer based on purchase history in Ecommerce market using clustering techniques. Thus it will be helpful for organization to identify the potential customer as well as apply marketing plans according to different segment of customers.Thus,identifying the right customer will help the enterprises to increase their profit and ultimately result into better customer relationship management.

Thus it will be helpful for enterprises to better identify their customer and apply different marketing strategy for group of customer and maintain customer satisfication. It will also helpful to customer since they get the services that they needed.

LIST OF FIGURES:

LIST OF TABLES:

TABLE OF CONTENTS

CHAPTERS **PAGE NO.**

Chapter 1

Introduction of Project

1.1 Introduction of Problem

The information plays a very important role in these competitive marketing. Every organization focuses on developing better customer relationship and customer satisfaction face of product competition, the organization should not only consider ever-changing market but also mine the customer resources to achieve targeted measures. Thus, the key to organization development is to start analysis of customer needs and perform customer segmentation on available data.

However, the traditional method of customer segmentation lags appropriate segmentation of customers. Itresults into poor clustering results and efficiency. Thus, a new clustering algorithm must be defined for efficient clustering of customer in order to increase the profit/revenue of organization and maintain customer relationship for longer time.

In our proposed system, we will combine two algorithms for providing better efficiency and clustering results. SAPK+K-Means are clustering technique that we are going to use for customer segmentation. It is also called as improved K-Means algorithm. Thus, combining both algorithms will outperform individual algorithm.

1.2 Problem statement and explanation

1.2.1 Problem statement

Customers are becoming more sophisticated in how they navigate their shopping choices and it is difficult to predict the customer purchasing pattern.It is necessary for everyorganization to identify the potential customers for maintaining good customer relationship as well as apply different marketing strategies for retaining their customers. It is necessary for enterprises to segment the different types of customer according to their value.

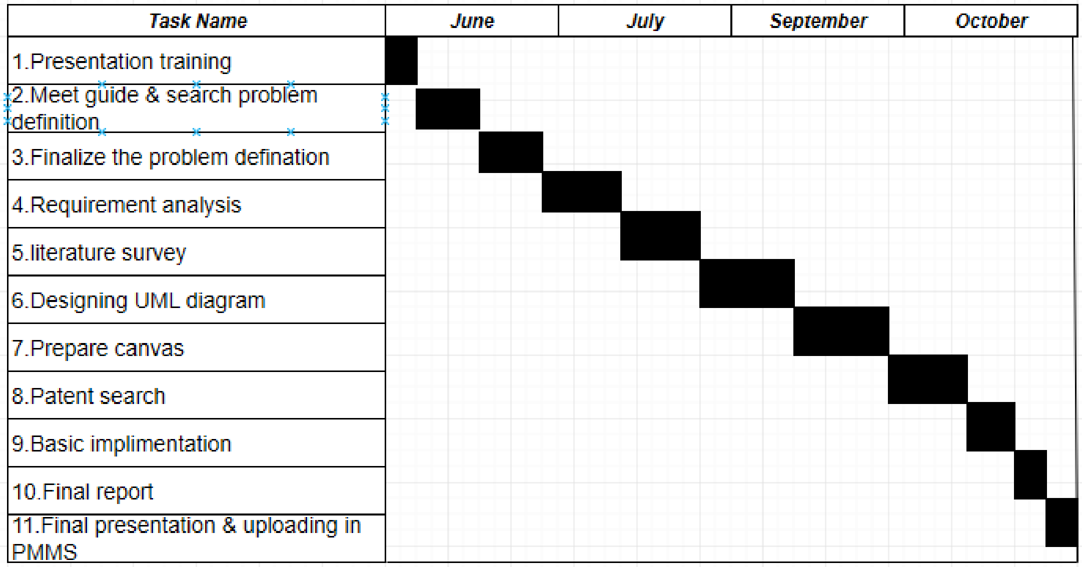
The traditional methods for customer segmentation lags proper segmentation of customer and is not efficient with complex data. Thus, it is not efficient way to obtain better clustering results.Moreover, targeting the wrong customer would result into not just in wasting the money for marketing as well as result into higher operational cost.Therefore,more efficient way is needed that would result into better clustering result and help in maintaining better customer relationship.

1.2.2 Description

In our proposed system, the dataset will be given as input to Semi Supervised affinity propagation(SAPK).SAPK algorithm is based on message passing between data points and it does not take initial cluster center or number of clusters. Thus, every data point will be considered as potential cluster at initial stage in SAPK.SAPK algorithm is used to find the similarity between data points and find the number of cluster and initial centroid until the convergence.

This number of cluster and cluster centroid generated from SAPK algorithm is given as input to K-Means algorithm. InK-Means algorithm, we need to specify the number of clusters and cluster centroidfor clustering. Run the K-Means algorithm to get clustering result. After getting clustering result, we can evaluate clustering result to identify characteristics of each clusters. Thus, it will provide better clustering result as compared to traditional methods and help the organization to target the potential customers with different marketing strategies.

1.3 Project Plan



(Time-line chart for phase-I)

Chapter 2

Literature Review

2.1 Existing Systems:

Systems uses many algorithms as an integral part for segmenting the customer to provide better clustering result. Every algorithm will have their own advantage and disadvantage for clustering. Algorithm like K-Means, Hierarchical and Affinity propagation has been discussed along with their features based on our survey.

2.1.1 Customer segmentation using Machine Learning:

- Benefits of our proposed system is that we do not have to select the initial number of cluster and cluster centroid at initial stage. Rather the system will provide such information after applying Semi-Supervised Affinity Propagation algorithm.
- It will provide better clustering quality and efficiency.
- Help the Organization to increase their profit and maintain better customer relationship and customer retention through promotional offers.

2.2Comparative statements from survey:

Features	K-Means Clustering	Hierarchical Clustering	Semi Supervised Affinity Propagation(SAPK)
Computational Speed	High	Low	Low
Clustering Time	Less	More	More
Required initial number of clusters	Yes	No	No
Capability to handle large datasets	Yes	No	Yes
Error Rate	More	More	Less
Clustering result efficiency	Low	Low	High

Table. 2.2 Comparative statement from survey

2.3 Motivation from survey:

By a conducted survey, we found that every organization focus on their customer needs in order to generate their revenue. There are traditional methods like K-Means available but it does provide proper clustering result. If the organization want to move ahead of its competitors then besides ever-changing market it should also analyze their customer history data to get meaningful insight that would help them to increase their revenue and customer retention through marketing strategies.

So, after referring the survey, we found that combining the two algorithms in order to generate better clustering result and efficiency as compared to traditional method is better options instead of individual algorithm. It will ultimately help the organization to mine the customer data and provide better customer services that the customer needs.

Chapter 3

System Analysis

3.1 Introduction

Customer Segmentation using Machine Learning is amachine learning based system, which help the organization to mine the customer past history data and efficiently segment the customer based on some characteristics between them. It will allow the organization to apply different marketing strategies based on the grouped data of customer. Itresults into increasing the productivity and profitability of the organization and maintain customer relationship.

3.1.1 Purpose

With the help of **Customer Segmentation using Machine Learning**organization can get idea about the types of customer exist in the current market. Efficient segmentation of customer to provide better services to customer, profit of organization and maintaining customer satisfaction is the main goal of our proposed system.

Providing the services that customer wants is the main motive of our proposed system to maintain customer retention.

3.1.2 Scope

With the help of Customer Segmentation using Machine Learning, organization can get an idea about who are the potential customers, basiccustomers, ordinarycustomers, excellent customers etc. and apply the different marketing strategies for different class of customer.

3.1.3 Document Conventions

Throughout this documentation, the following conventions have been used:

- Fonts: Times New Roman
- Size 16 for Main Headings
- Size 14 for Sub Headings
- Size 12 for the Rest of the Document

3.2 Overall Description

3.2.1 Product Perspective

Customer Segmentation using Machine Learning is organization-based system that follows Customer-Organization interaction. It will help the organization to make better decision about the different class of customers.

3.2.2 Product Function

Customer Segmentation using Machine Learning has the provide the following functions:

- Analyze the input data.
- Remove noisy data, outliers etc.
- Apply algorithm to generate number of clusters.
- Clusters based on similarity.
- Output the clustering results.
- Describe the characteristics of each clusters.

3.2.3 User Characteristics

Every Organization have the responsibility to analyze the purchase history data of each customers. They should know how to cluster the given data for getting meaningful insight of that data. They should try to improve its efficiency and clustering results.

3.2.4 Constraints, Assumptions and Dependencies

1) **Hardware Limitations**

- A computer with a web browser installed - preferably latest versions of Chrome, Firefox or Internet Explorer
- A computer should have an active internet connection.

2) **Safety and Security considerations**

Since everything is based on dataset in our proposed system it may possible that the data, we gather may contain noise,outliers,missing values etc. In such case we have to first perform data preprocessing before we apply algorithm on given dataset to cluster the customer data.

3) Criticality of Application

If the given input dataset is very large, it may possible that the system takes more time for generating clustering results.

4) Assumptions and Dependencies

Our proposed system highly depends on the given dataset as input for clustering. Better and more specific the given dataset, the better clustering result can be obtained.

3.3 Specific Requirements

3.3.1 External Interface Requirements

3.3.1.1User Interface

Better screen resolution will be helpful for getting better visualization of clusters for given input dataset.

3.3.1.2 Hardware Interfaces

- PC/Laptop
- Internet connectivity

3.3.1.3 Software Interfaces

- Operating System(Windows 7 or above)
- Any web browser - Chrome is more preferred.
- Python 2.7 or above
- Jupiter Notebook

3.3.1.4 Communication Interface

- Python 2.7 or above is must.

3.3.2 Functional Requirement

3.3.2.1Customer/Market Segmentation

Market segmentation is the research that determines how your organization divides its customers into smaller groups based on characteristics such as, age, income, personality traits or behavior.

3.3.2.2Measurable

The size, purchasing power, and profiles of the segments can be measured.

3.3.2.3 Data Manipulation

Data manipulation is the process of changing data to make it easier to read or be more organized.

3.3.2.4 Data Processing

It is technique that involves transforming raw data into an understandable format.

3.3.2.5Data Analysis

Data analysis is the process of evaluating data using analytical and statistical tools to discover useful information and aid in business decision making.

3.3.3 Other Non-functional Requirements

3.3.3.1 Accessible

The market segments must be effectively reached and served.

3.3.3.2 Substantial

Something substantial is large in size, number, or amount. If you want to say someone spent a lot of money without being too specific, you could say they spent a substantial amount of money.

3.3.3.3 Accuracy

Accuracy is the measurement used to identifying relationships and patterns between variables in a dataset based on the input, or training data.

3.3.3.4 Clearly Defined and Distinguishable Segment

The chosen segments should be clearly defined to avoid doubt about which part of market.

Chapter 4
System Design

4.1 System Flow Chart

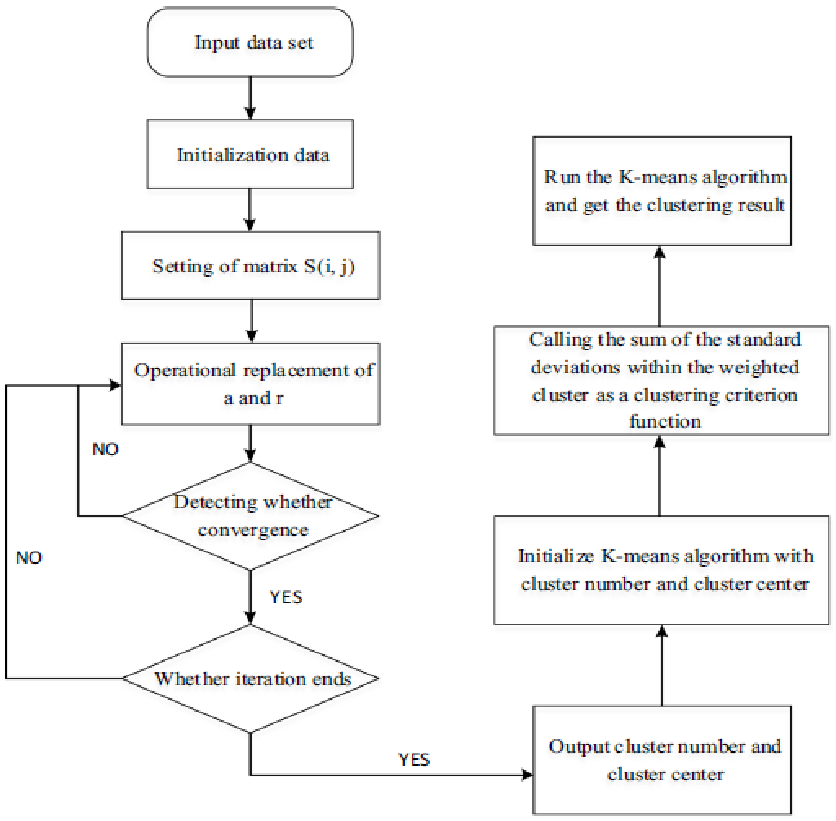

Input data set
Initialization data
Setting of matrix S(i, j)
Operational replacement of a and r
NO
Detecting whether convergence
NO
YES
Whether iteration ends
YES
Output cluster number and cluster center
Initialize K-means algorithm with cluster number and cluster center
Calling the sum of the standard deviations within the weighted cluster as a clustering criterion function
Run the K-means algorithm and get the clustering result

Fig. 4.1 System Flowchart

4.2 UML Diagrams

The **Unified Modelling Language**(**UML**) is a standard visual modelling language intended to be used for

- modelling business and similar processes,
- analysis, design, and implementation of software-based systems

UML is a common language for business analysts, software architects and developers used to describe, specify, design, and document existing or new business processes, structure and behaviour of artefacts of software systems.

4.2.1 Use case Diagram

Use case diagrams model the functionality of a system using actors and use cases. Use cases are services or functions provided by the system to its users.

So, in a brief when we are planning to draw a use case diagram, we should have the following items identified.

- Functionalities to be represented as a use case
- Actors
- Relationships among the use cases and actors.

Use case Diagram of System:

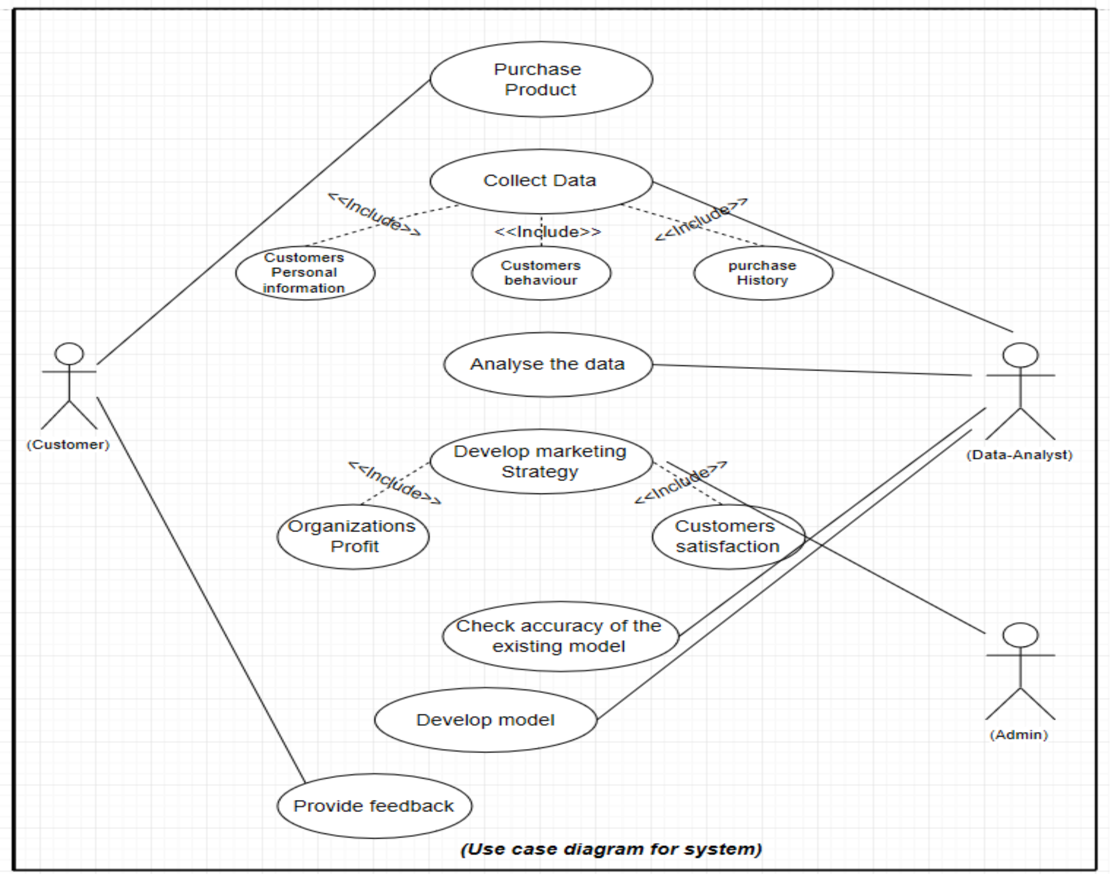
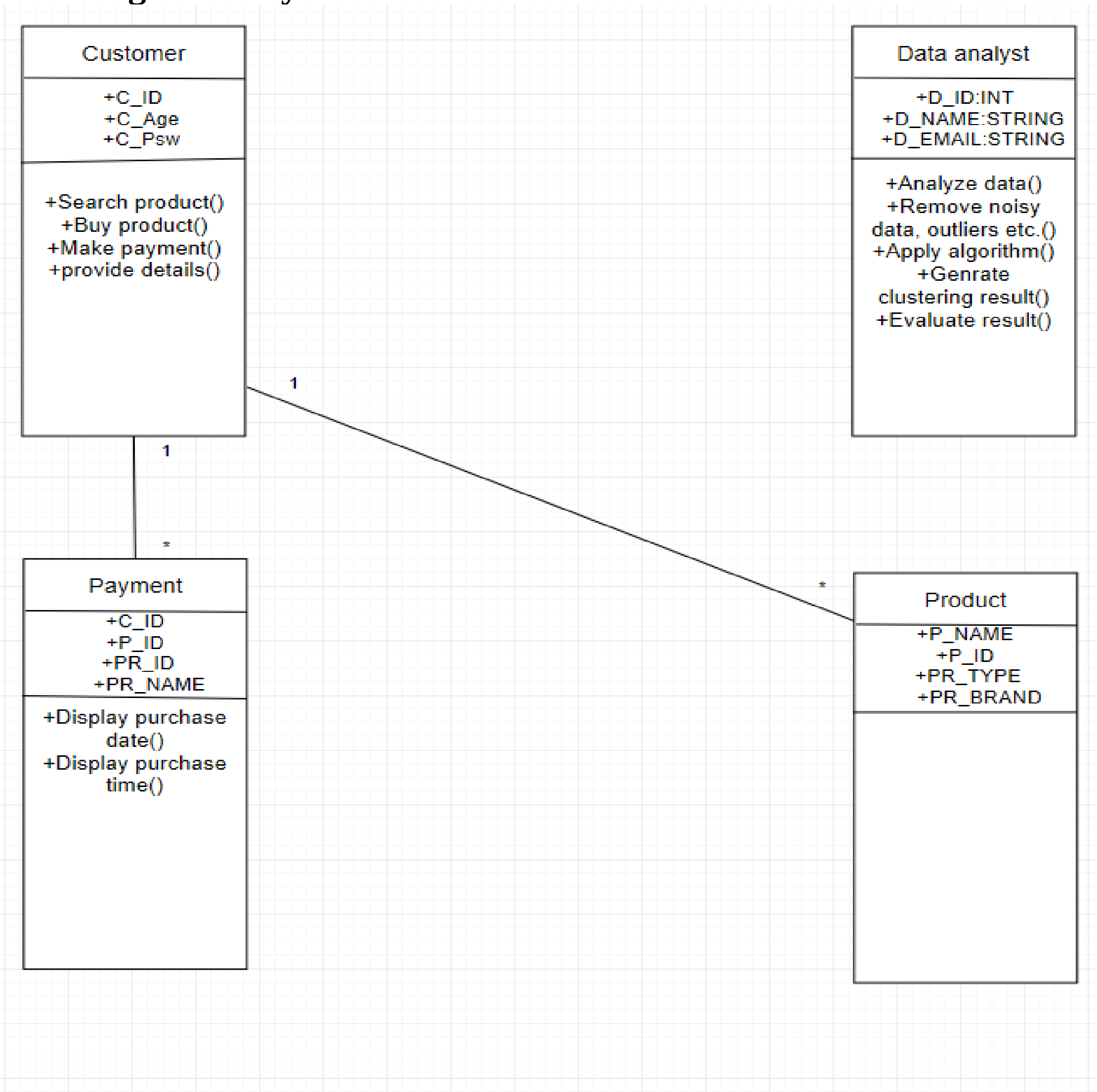


Fig. 4.2.1 Use case Diagram of system

4.2.2 Class Diagram

Class diagrams are the backbone of almost every object-oriented method including UML. They describe the static structure of a system. In software engineering, a **class diagram** in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

Class Diagram of System:



(Class diagram for system)

Fig. 4.2.2 Class Diagram of System

4.2.3 Sequence Diagram

Sequence diagrams describe interactions among classes in terms of an exchange of messages over time.

A sequence diagram is an interaction diagram. From the name it is clear that the diagram deals with some sequences, which are the sequence of messages flowing from one object to another.So, Sequence diagram is used to visualize the sequence of calls in a system to perform a specific functionality.

Sequence Diagram of System:

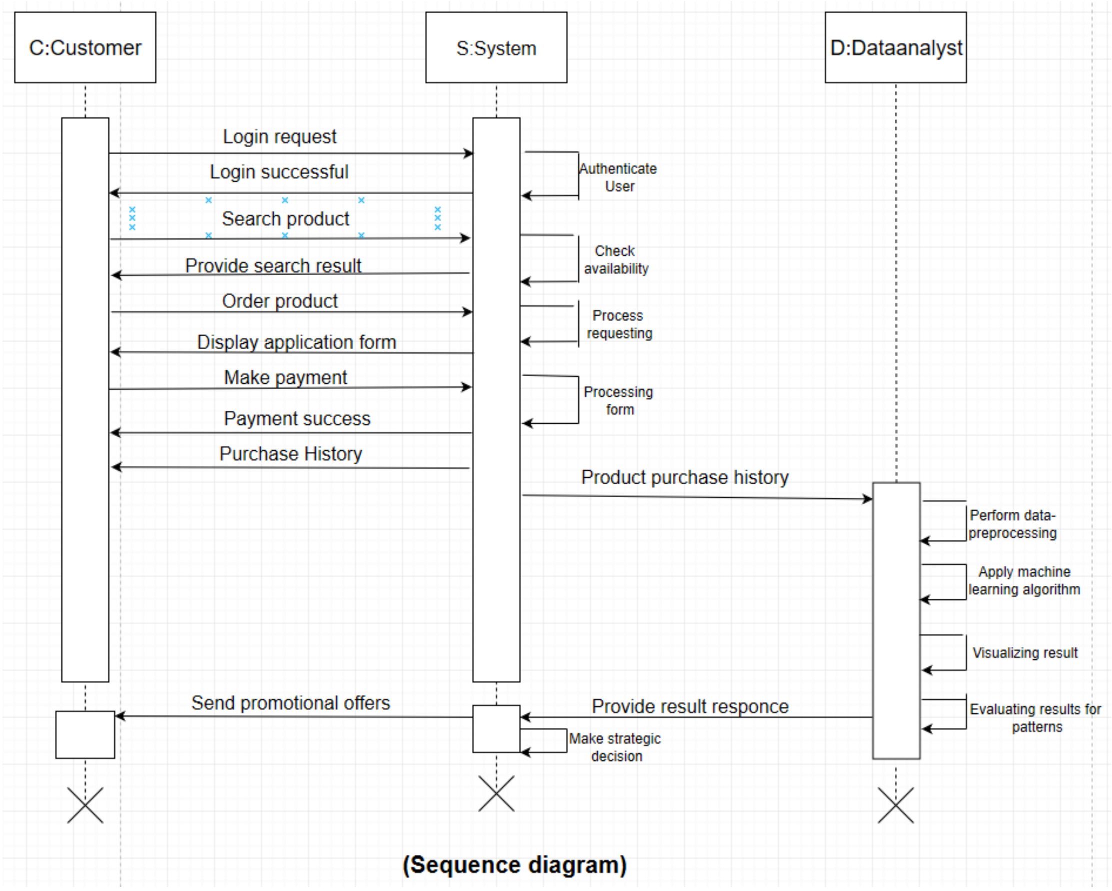


Fig. 4.2.3 Sequence Diagram of System

4.2.4 Activity Diagram

An activity diagram illustrates the dynamic nature of a system by modeling the flow of control from activity to activity. An activity represents an operation on some class in the system that results in a change in the state of the system.

Activity Diagram of system:

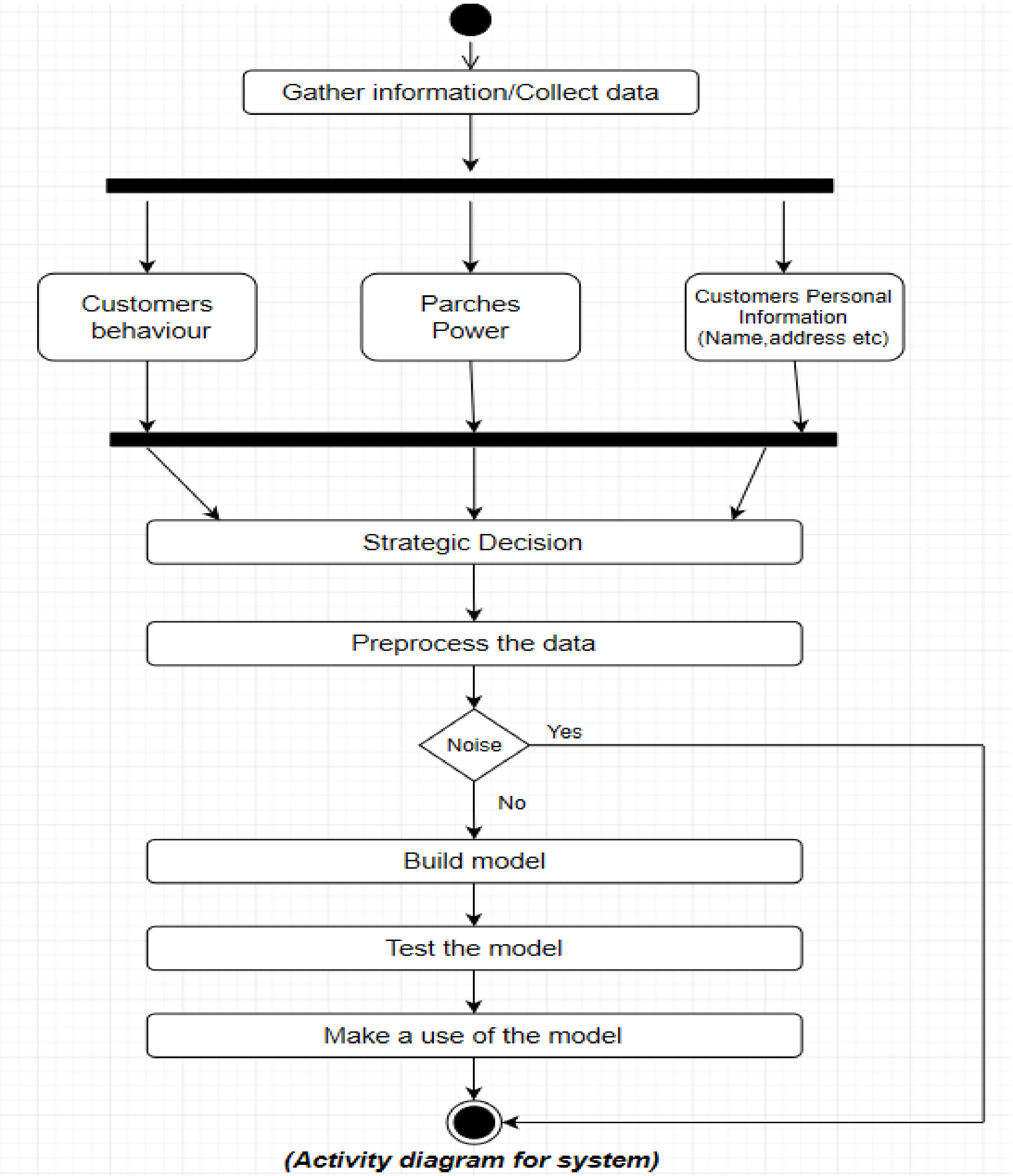


Fig. 4.2.4 Activity Diagram of System

4.2.5 State Chart Diagram

A State chart diagram describes a state machine. Now to clarify it state machine can be defined as a machine which defines different states of an object and these states are controlled by external or internal events.

State chart diagram describes the flow of control from one state to another state. States are defined as a condition in which an object exists and it changes when some event is triggered. So, the most important purpose of State chart diagram is to model life time of an object from creation to termination.

State Chart Diagram of system:

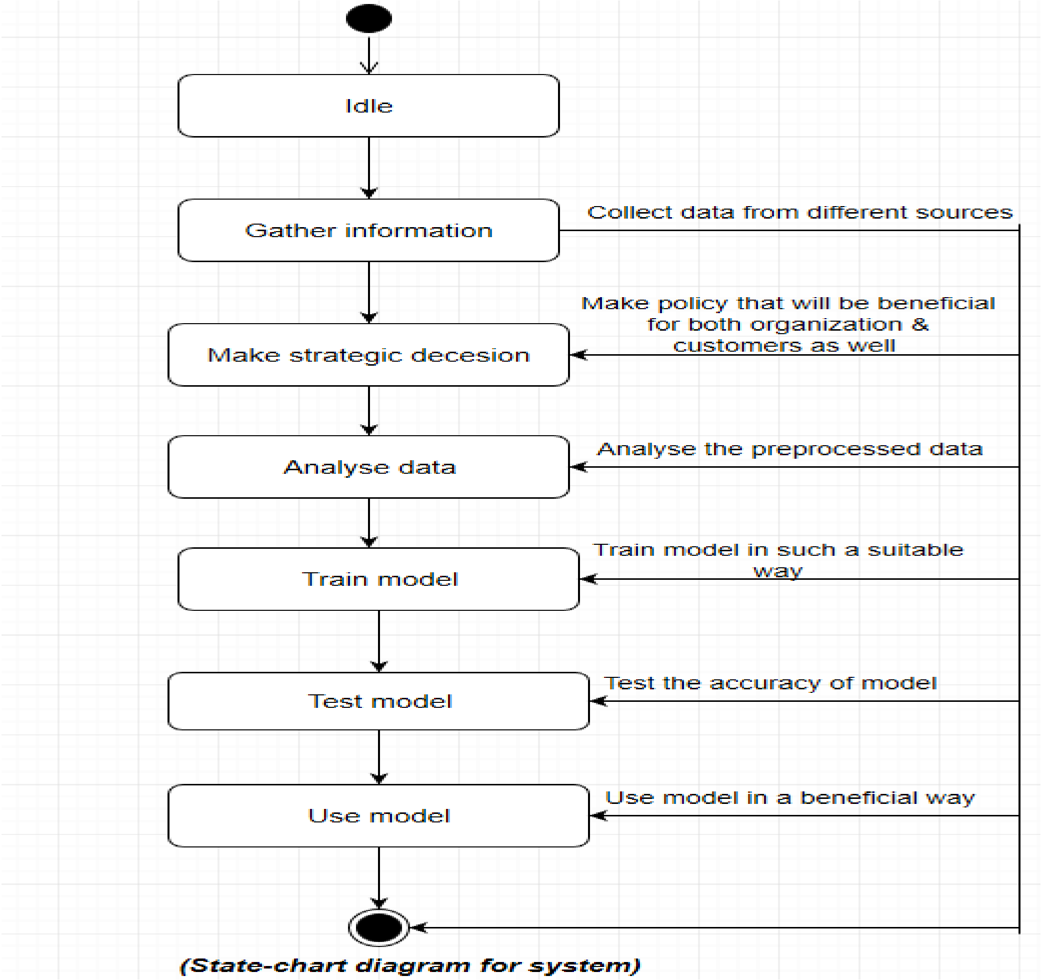

Idle
Gather information
Collect data from different sources
Make policy that will be beneficial for both organization & customers as well
Make strategic decesion
Analyse the preprocessed data
Analyse data
Train model in such a suitable way
Train model
Test the accuracy of model
Test model
Use model in a beneficial way
Use model
(State-chart diagram for system)

Fig. 4.2.5 State Chart Diagram of System

4.2.6 Collaboration Diagram

A Collaboration diagram is very similar to a Sequence diagram in the purpose it achieves; in other words, it shows the dynamic interaction of the objects in a system. A distinguishing feature of a Collaboration diagram is that it shows the objects and their association with other objects in the system apart from how they interact with each other. The association between objects is not represented in a Sequence diagram.

Collaboration Diagram of system:

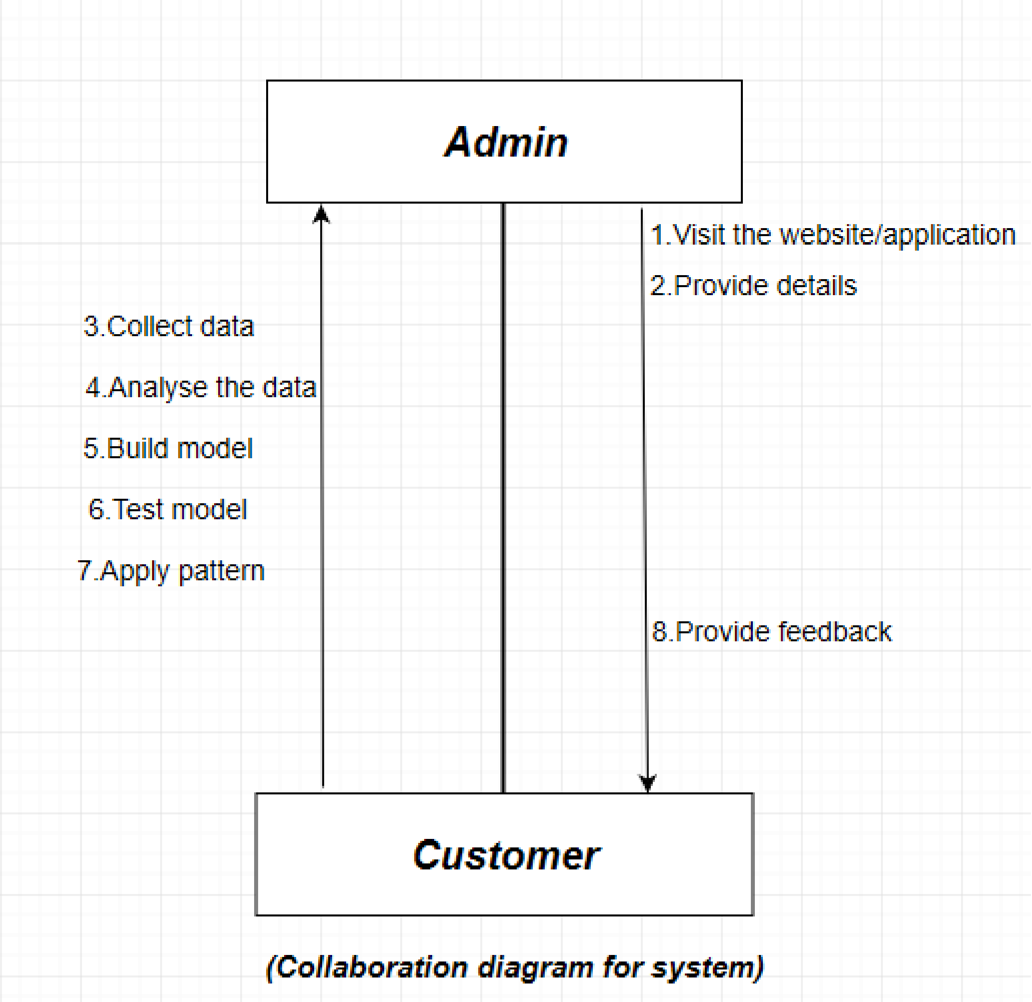


Fig. 4.2.6 Collaboration Diagram of System

4.2.7 Component Diagram

Component diagram is a special kind of diagram in UML. The purpose is also different from all other diagrams discussed so far. It does not describe the functionality of the system but it describes the components used to make those functionalities.

So, from that point component diagrams are used to visualize the physical components in a system. These components are libraries, packages, files etc.

Component Diagram:

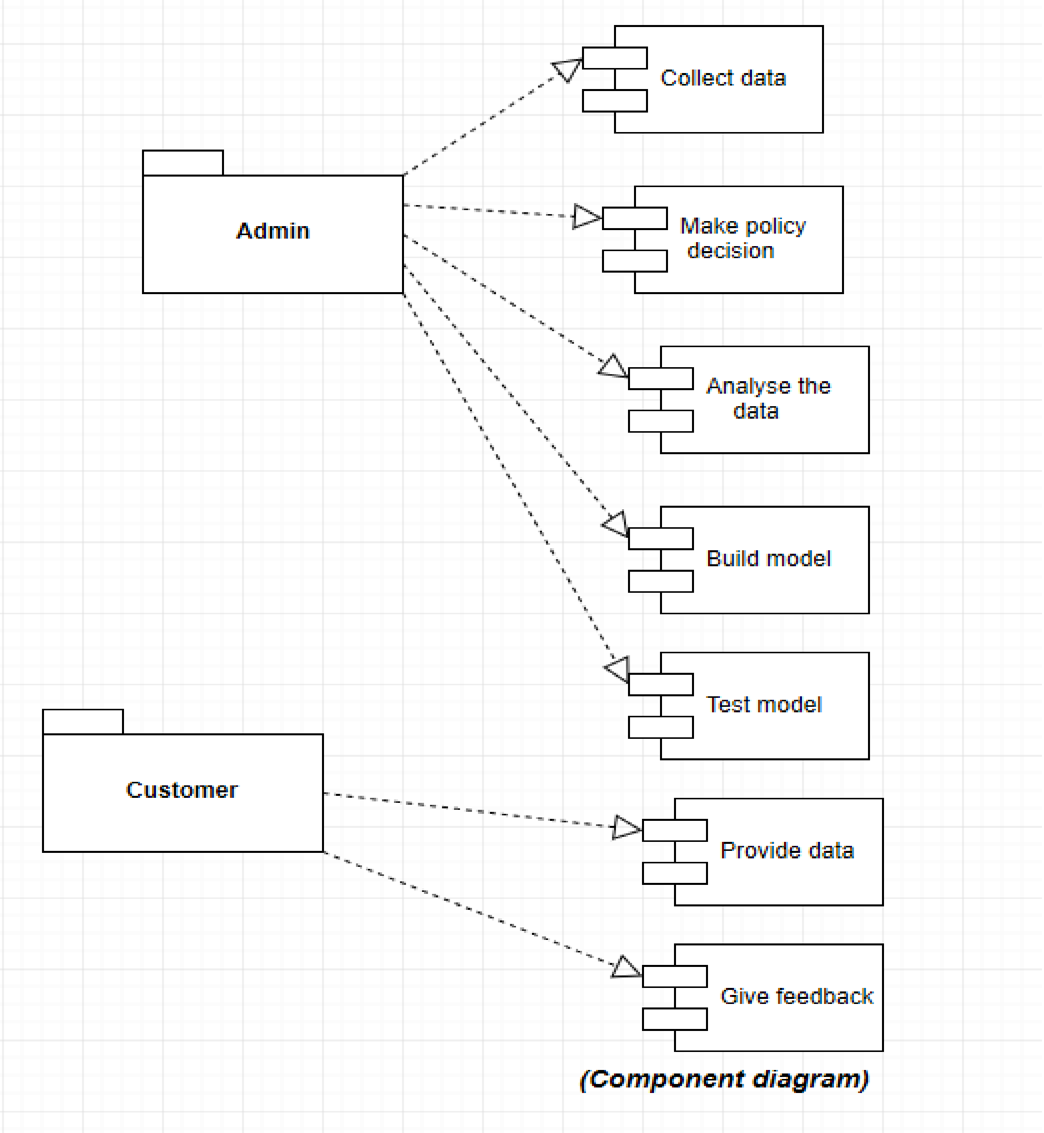


Fig. 4.2.7 Component Diagram of System

4.3 Database Design

4.3.1 Entity Relationship Diagrams

Entity-Relationship model making possibility to describe a database in which in the table's data can be the point to data in other tables - for instance, your entry in the database could point to several entries.

ER Diagram of System:

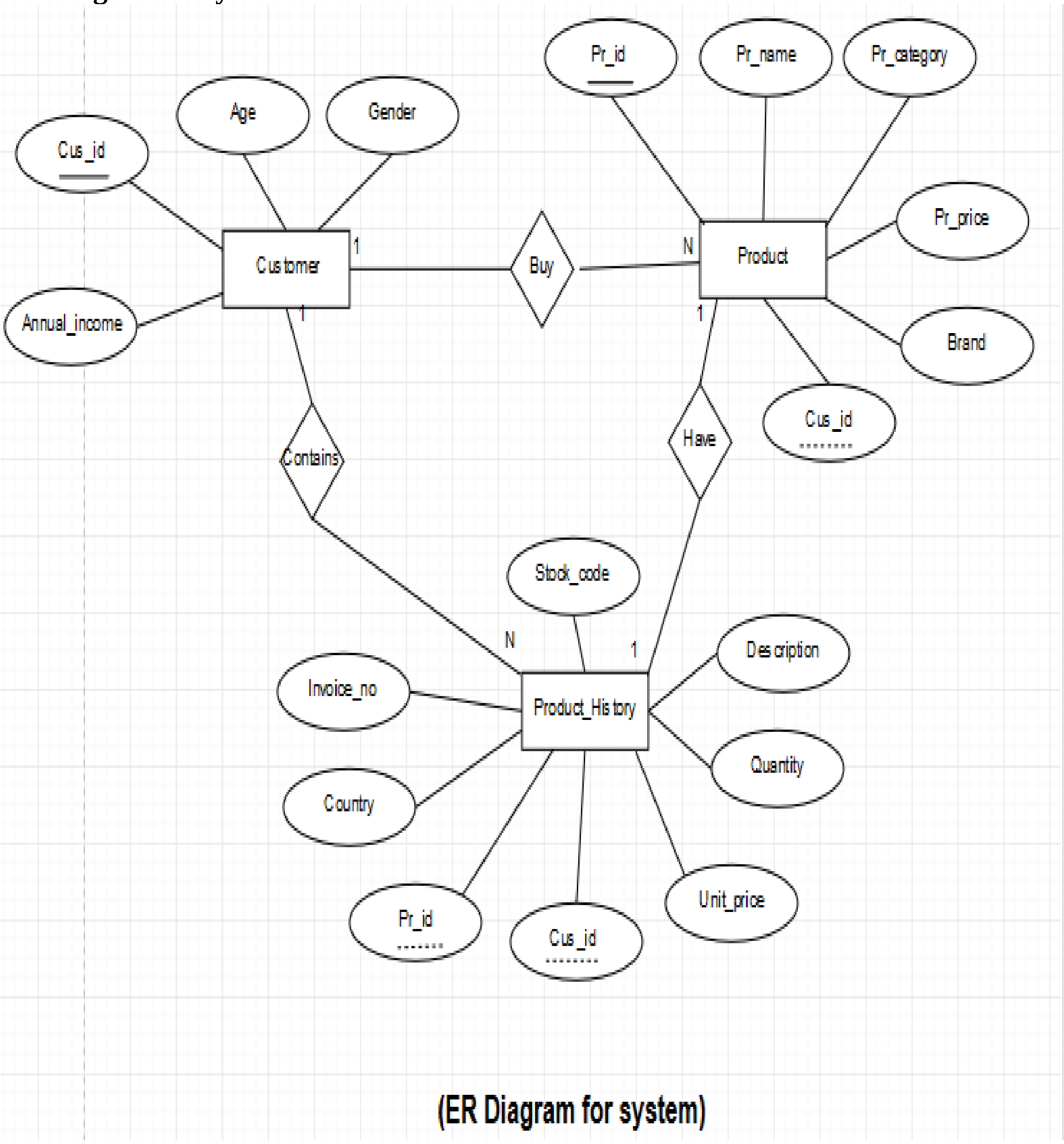


Fig.4.3.1 ER Diagram of system

4.3.2 Data Dictionary:

1)For Customer:

Field	Data Type	Size	Description
Cus_id(PK)	varchar	(20)	CustomerId
Cus_age	int	(10)	CustomerAge
Cus_gender	varchar	(20)	CustomerGender
Cus_income	float	(20)	CustomerIncome

Table. 4.3.2.1 Customer

2) For Product:

Field	Type	Size	Description
Pr_id(PK)	varchar	(20)	ProductId
Pr_name	varchar	(20)	ProductName
Pr_category	varchar	(20)	ProductCategory
Pr_price	float	(10)	ProductPrice
Pr_brand	varchar	(20)	ProductBrand
Cus_id(FK)	varchar	(20)	CustomerId

Table. 4.3.2.2Product

3)Product_history:

Field	Type	Size	Description
Invoice	int	(11)	ProductInvoice
Stockcode	varchar	(20)	ProductStockCode
Description	varchar	(100)	ProductDescription
Quantity	int	(10)	ProductQuantity
UnitPrice	float	(10)	ProductUnitPrice
Cus_id(FK)	varchar	(20)	CustomerId
Pr_id(FK)	varchar	(20)	ProductId

Table. 4.3.2.3Product_history

4.3.3 Data Flow Diagram:

Data flow diagrams present the logical flow of information through a system in graphical or pictorial form. Data flow diagrams have only four symbols, which makes useful for communication between analysts and users. Data flow diagrams (DFDs) show the data used and provided by processes within a system.

DFD Level 0:

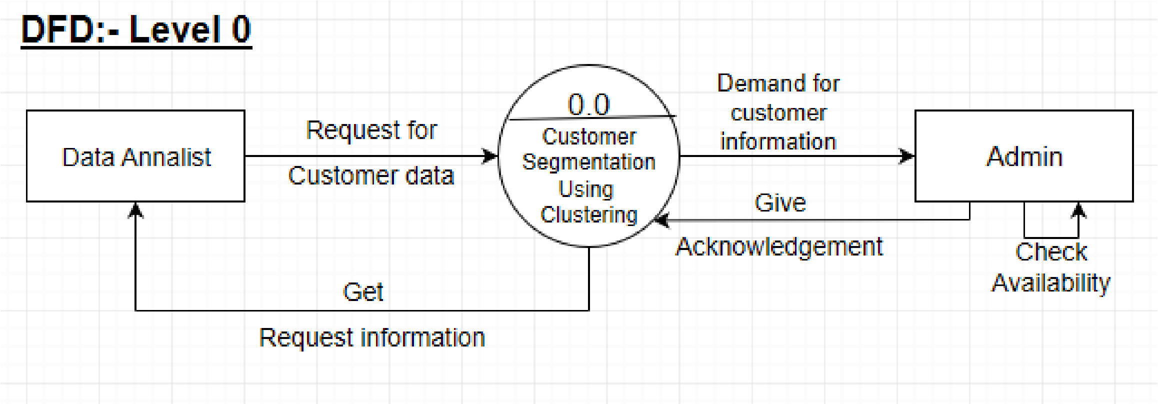


Fig. 4.3.3.1 Level 0 DFD

DFD Level 1 (Admin):

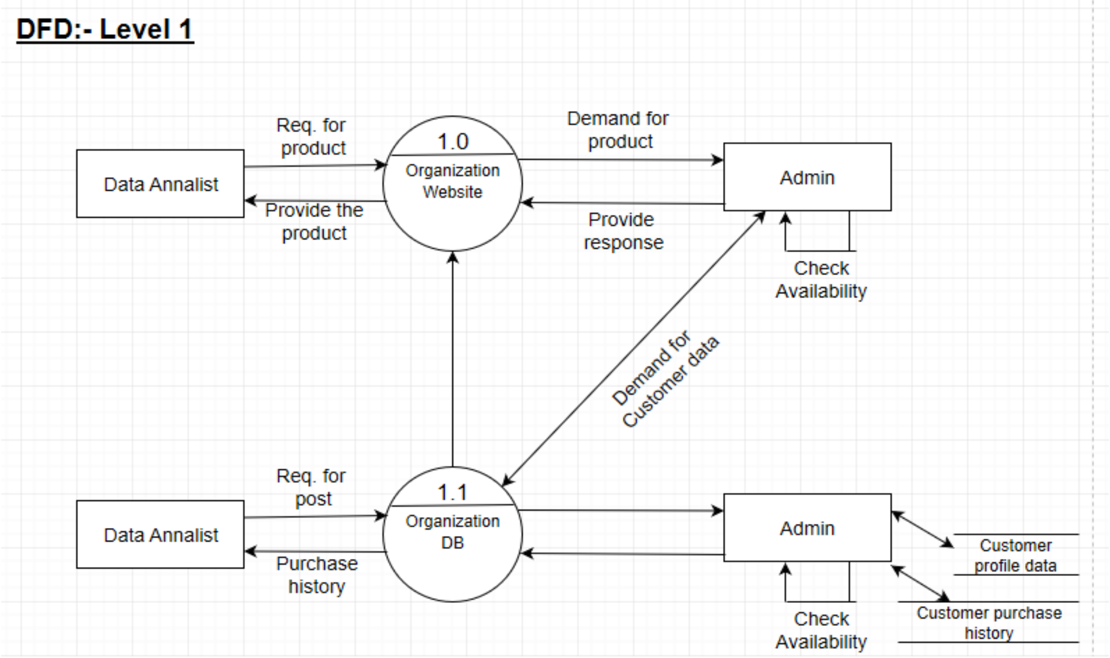


Fig. 4.3.3.2 Level 1 DFD (Admin)

DFD Level 2 :

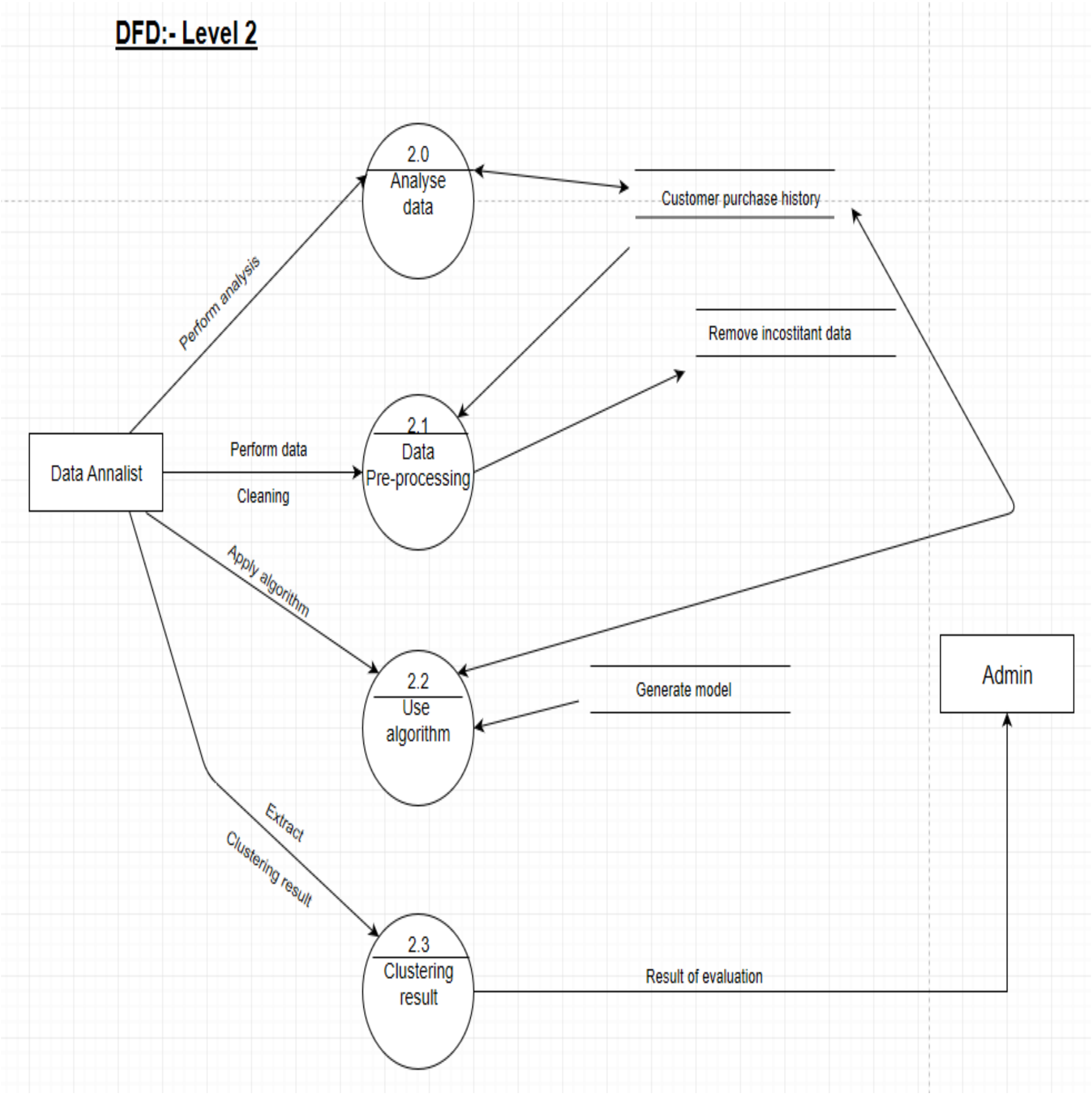


Fig.4.3.3.3 Level 2 DFD

Chapter 5

Canvas

5.1 AEIOU summary:

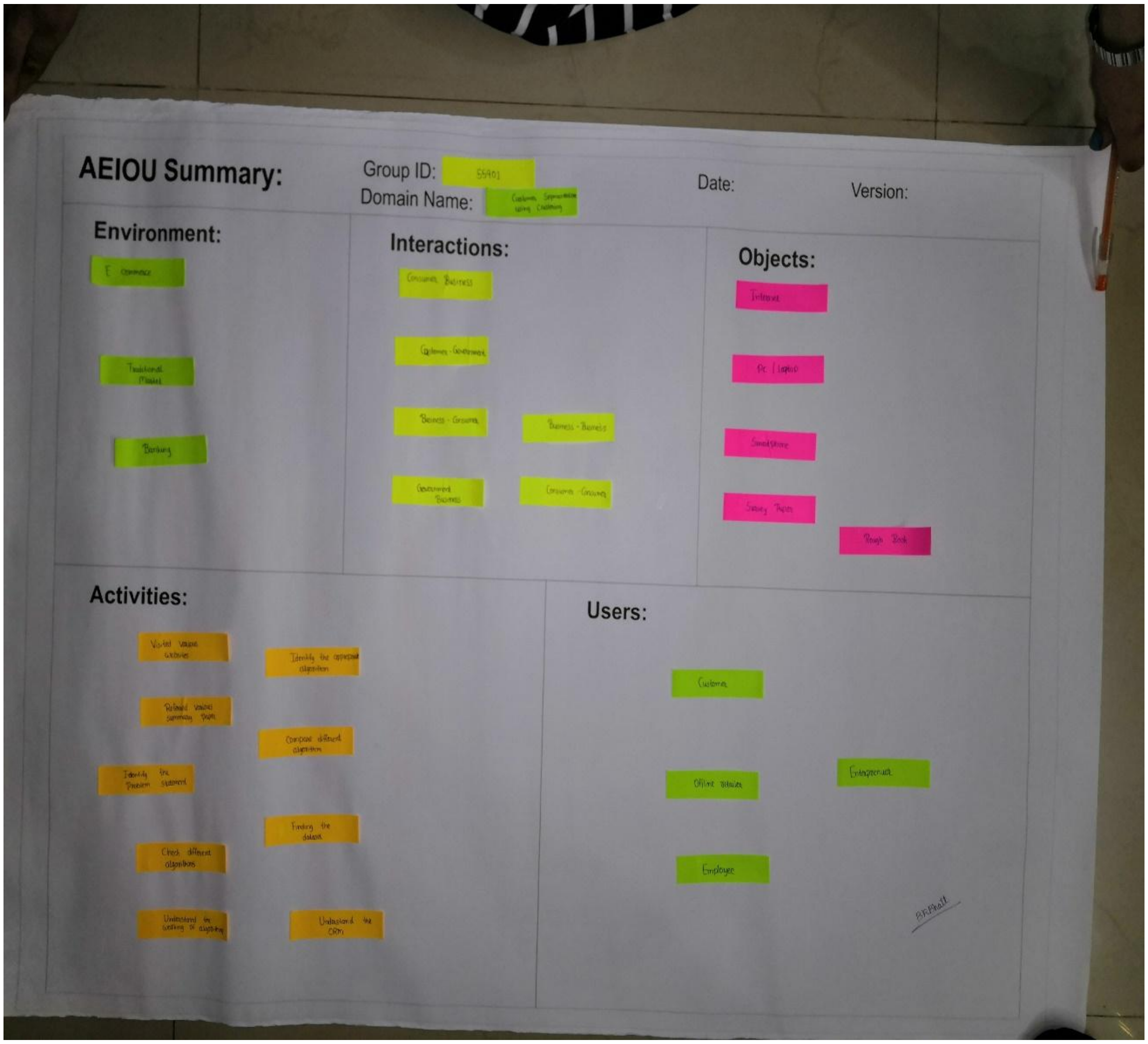
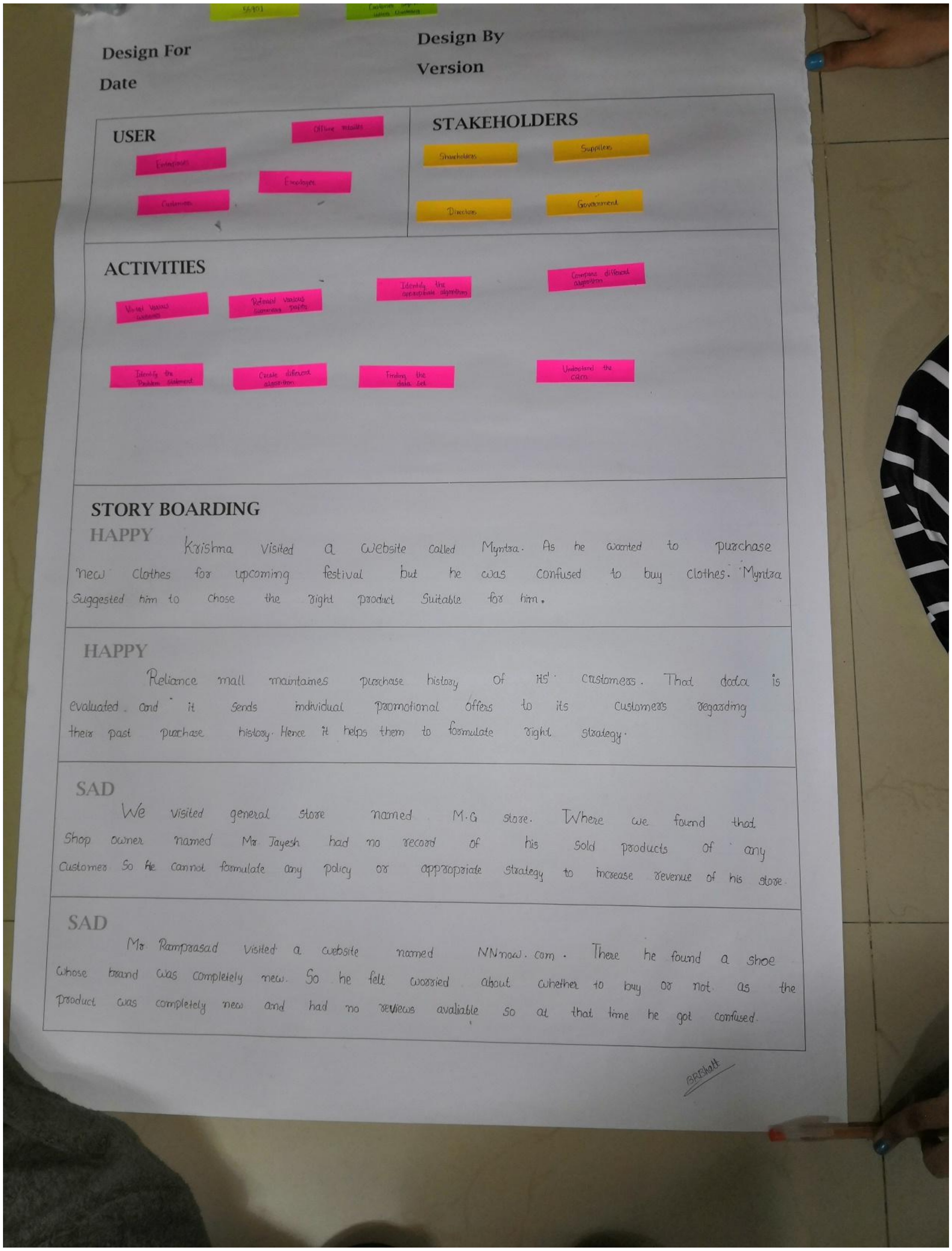


Fig.5.1 AEIOU Canvas

5.2 Empathy:



Design For

Date

Design By

Version

USER

Offline retailers

Enterprises

Employee

Customers

STAKEHOLDERS

Shareholders

Suppliers

Directors

Government

ACTIVITIES

Visited various websites

Referred various summaries papers

Identify the appropriate algorithm

Compare different algorithm

Identify the Problem statement

Create different algorithm

Finding the data set

Understand the CRM

STORY BOARDING

HAPPY

Krishna visited a website called Myntra. As he wanted to purchase new clothes for upcoming festival but he was confused to buy clothes. Myntra suggested him to chose the right product suitable for him.

HAPPY

Reliance mall maintaines purchase history of its' customers. That data is evaluated and it sends individual promotional offers to its customers regarding their past purchase history. Hence it helps them to formulate right strategy.

SAD

We visited general store named M.G store. Where we found that shop owner named Mr Jayesh had no record of his sold products of any customer so he cannot formulate any policy or appropriate strategy to increase revenue of his store.

SAD

Mr Ramprasad visited a website named NNnow.com. There he found a shoe whose brand was completely new. So he felt worried about whether to buy or not as the product was completely new and had no reviews avaliable so at that time he got confused.

Fig.5.2 Empathy Canvas

5.3 Ideation:

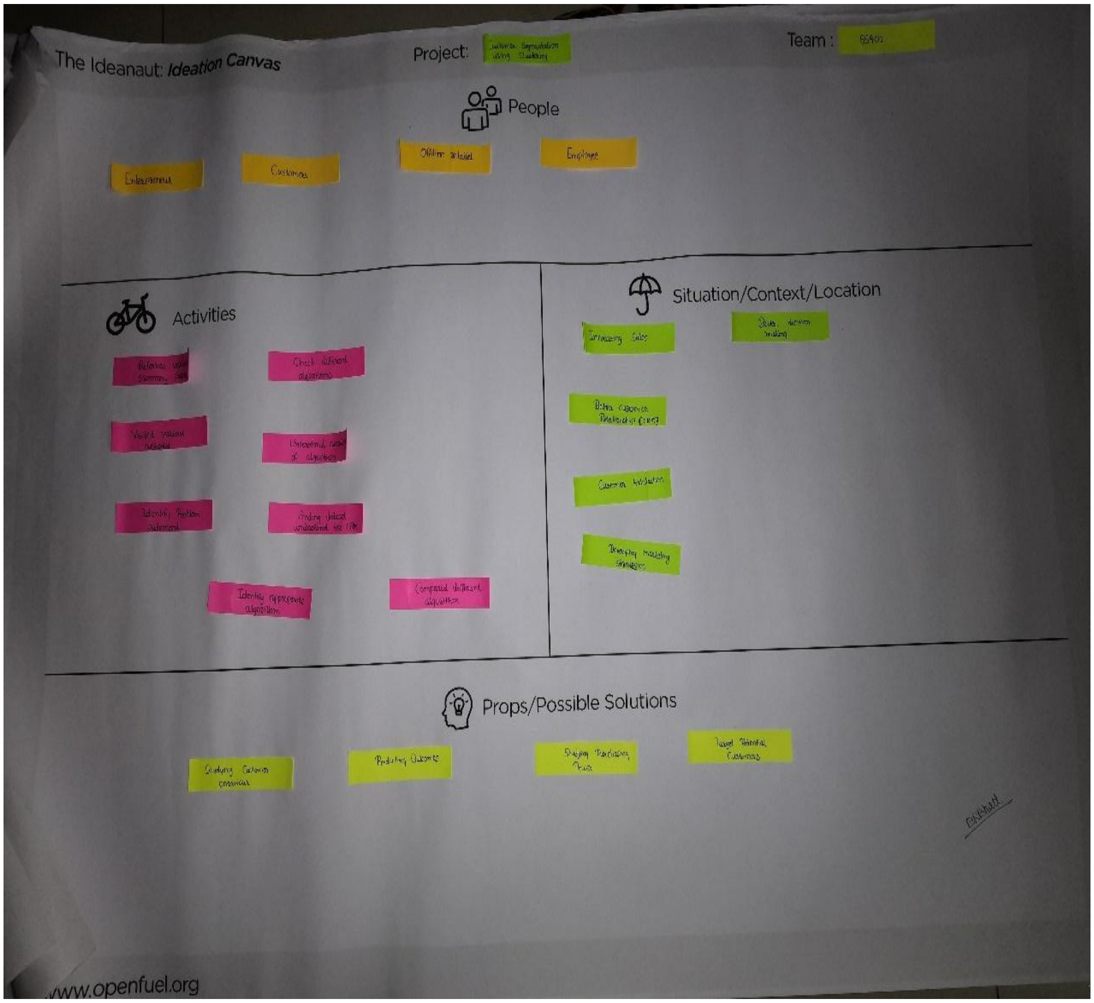

The Ideanaut: Ideation Canvas
Project: Customer Segmentation using Clustering
Team :
People
Entrepreneur
Customers
Offline retailer
Employee
Activities
Reference various summary paper
Check different algorithms
Visited various websites
Understand working of algorithm
Identify Problem statement
Finding dataset understand the CRM
Identify appropriate algorithm
Compared different algorithm
Situation/Context/Location
Increasing Sales
Data driven making
Better customer Relationship (CRM)
Customer Satisfaction
Developing marketing strategies
Props/Possible Solutions
Studying Customer behaviour
Predicting Outcomes
Shaping Purchasing Trends
Target Potential Customers
www.openfuel.org

Fig.5.3 Ideation Canvas

5.4 Product Development:

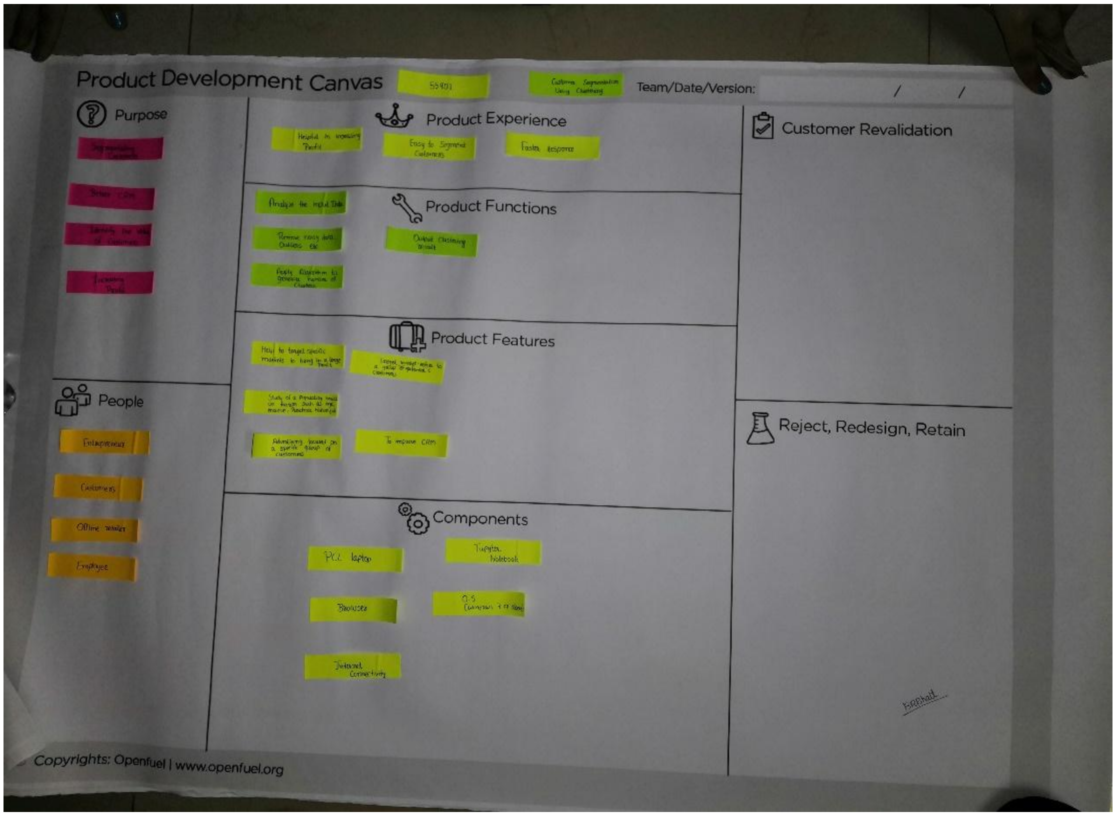


Fig.5.4 PDC Canvas

Chapter 6

Expected Outcome

Upon the completion of the project, our team would have built a product which will help for both organization as well as customer.It would be helpful for organization to efficiently identify potential customers.Customer get satisfied when they get services what they needed.

Following are the function that our proposed system will have:

Analyze the input data:

- Our proposed system will analyze the input dataset gather from different sources.

Remove noisy data,outliers etc:

- After analyzing the data,if there are noisy data,outliers,missing values etc then our proposed system will perform data preprocessing before using that data for clustering.

Apply the algorithm:

- After removing inconsistent data using data preprocessing,we will provide our data as input to the algorithm in order to generating clustering result.
- Clustering will be perform on the basis of some similarity between behavior of customer.

Evaluating the clustering result:

- After getting the clustering result,we can evaluate the number of clusters on basis on their characteristics.
- Thus,this clustering result will help the organization to identify value of their customers and apply different marketing strategies based on clustering result for each clusters.

Conclusion

By analyzing the deficiencies and existing defects of the traditional K-means algorithm and combining with AP algorithm, a new algorithm combining the semi-supervised APalgorithm and the classic K-means algorithm is proposed, namely SAPK + K-means algorithm.

The improved SAPK + K-means algorithm has a low error rate in the clusters but the clustering time is relatively long, which to some extent guarantees the high validity and accuracy of the SAPK + K-means algorithm for the clustering analysis of data information. Thus, by combining the traditional K-Means with Affinity Propagation better clustering results can be obtained which results into better understanding of customers characteristics and help in maintaining better customer relationship.

Hence, it would be helpful for the organization for identifying the potential customers that increases their profit.It also help them in maintaining customer relationship and customer retention by executing different marketing strategies.

References

[1]Deng, Yulin, and Qianying Gao. “A study on e-commerce customer segmentation management based on improved K-Means Algorithm”.Information System and E-Business Management(2018):1-14.

[2]Tripathi, S., A. Bhardwaj, and E. Poovammal. "Approaches to clustering in customer segmentation." *International Journal of Engineering & Technology* 7.3.12 (2018): 802-807.

[3]Morisada, Makoto, Yukihiro Miwa, and Wirawan Dony Dahana. "Identifying valuable customer segments in online fashion markets: An implication for customer tier programs." *Electronic Commerce Research and Applications* 33 (2019): 100822.

[4]Xiaoman, Wei, et al. "Analysis of power large user segmentation based on affinity propagation and K-means algorithm." *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*. IEEE, 2017.

[5] Bhade, Kalyani, et al. "A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization." *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2018.

[6]Huang, S. L., Q. Wang, and B. School. "Method for customer segmentation based on three-way decisions theory." *Journal of Computer Applications* 34.1 (2014): 244-248.

[7]Abdi, Farshid, and Shaghayegh Abolmakarem. "Customer Behavior Mining Framework (CBMF) using clustering and classification techniques." *Journal of Industrial Engineering International* (2018): 1-18.

[8]Ye, Luo, et al. "Customer segmentation for telecom with the k-means clustering method." *Information Technology Journal* 12.3 (2013): 409-413.

[9]https://www.datasciencecentral.com/profiles/blogs/the-best-ways-to-segment-customer-data