Developing a Machine Learning Model for Predicting Purchasing Behavior of F-Commerce-based Customer

by

Mehnaz Fatema

DEPARTMENT OF INDUSTRIAL AND PRODUCTION ENGINEERING (IPE)

BANGLADESH UNIVERSITY OF ENGINEERING & TECHNOLOGY (BUET)

12th November, 2022

Developing a Machine Learning Model for Predicting Purchasing Behavior of F-Commerce-based Customer

by

Mehnaz Fatema

A project submitted to the Department of Industrial & Production Engineering, Bangladesh University of Engineering and Technology, in the partial fulfilment requirements for the degree of Master of Engineering in Advanced Engineering Management.

DEPARTMENT OF INDUSTRIAL AND PRODUCTION ENGINEERING (IPE)

BANGLADESH UNIVERSITY OF ENGINEERING & TECHNOLOGY (BUET)

12th November, 2022

CERTIFICATE OF APPROVAL

The project titled "Developing a Machine Learning Model for Predicting Purchasing Behaviour of F-Commerce-based Customer" submitted by Mehnaz Fatema, Student no: 1018082122 has been accepted as satisfactory in partial fulfillment of the requirements for the degree of **Masters of Engineering in Advanced Engineering Management** on 12th November, 2022.

BOARD OF EXAMINERS

Dr. Ferdous Sarwar
Professor
Department of IPE, BUET, Dhaka

Chairman

Dr. Syed Mithun Ali
Professor
Department of IPE, BUET, Dhaka

Member

Dr. Shuva Ghosh
Associate Professor
Department of IPE, BUET, Dhaka

Member

CANDIDATE'S DECLARATION

It is hereby declared that this project or any part of it has not been submitted elsewhere for the award of any degree or diploma.

_____

Mehnaz Fatema

Student no: 1018082122

DEDICATION

To my loving parents, my supportive junior and friend for their constant support throughout my career.

ACKNOWLEDGMENT

I want to express my deep appreciation to Dr. Ferdous Sarwar, the Head of the Industrial and Production Engineering Department at Bangladesh University of Engineering and Technology (BUET), for his unwavering support and guidance throughout the project. Dr. Sarwar provided me with his valuable professional advice and direction at every stage of the project, enabling me to develop as an independent researcher in the field of F-commerce. It was an honor to be part of his research team, and I feel fortunate to have had him as my supervisor for my master's thesis. Finally, I would be grateful to Almighty for providing me with the guidance and strength that helped me achieve this success. One part of my gratefulness also goes to the unrelenting support and encouragement of my family, friends, and peers.

ABSTRACT

The modern tech-based lifestyle has become much easier with online purchases. Social media platforms like Facebook, Instagram and Twitter have taken it to another level. Online business through Facebook is now a significant topic worldwide, so are the challenges. With the easy access of internet and free use of Facebook, doing business through Facebook platform is not a big deal. But as anyone, anywhere can start a small business with little capital using this giant platform due to its easiness, the competition and long-time survival is also the biggest of all challenges. Hence the aim of this study is to build a model that would help anyone get acknowledged with the satisfaction level of F-commerce users and target those pleased consumers in the future to serve better and minimize the survival challenges. The objective of this project is to predict the F-commerce based customer's satisfactory level and next preferable purchase. In this study, a survey method has been used to get real time customer data and a machine-learning model has been made that would help to cluster those customers together who prefer to purchase similar categories of product from Facebook. This project then also aims to classify and predict the satisfaction level of those customers in their purchase behaviour with real-world data by using K-medoid with the PCA (Principal Component Analysis) algorithm and SVM (Support Vector Machine).

Table of Content

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Introduction

Nowadays, it's rare to come across someone who is unfamiliar with the concept of electronic commerce. E-commerce refers to the practice of purchasing goods or services online, directly from vendors via the Internet. Customers can easily browse and find products they're interested in by visiting a retailer's website or using a shopping search engine to compare and contrast different sellers.

When such online activities are done by using the Facebook app it is called F-commerce. It mainly refers to using a Facebook page or Facebook application to conduct online business. F-commerce also refers to customers sharing information about purchases on Facebook, whether they use the Facebook "Share" tool on an e-commerce site or share information about purchases made at a physical point of sale. F-commerce is based on the idea of introducing a new type of online shopping: social shopping.

As of 2021 there are 2.85 billion active Facebook users (monthly) worldwide. This equates to almost one-third of the world's total population. Almost a billion of them are frequent visitors, logging in at least twice a week. And In Bangladesh, there are 44.7 million Facebook users as of January 2022 (S.Dixon, 2022). The platform's large user base has quickly inspired small enterprises to use it as a marketplace. There are over 2000 of dedicated e-commerce platforms and over 300,000 F-commerce pages in operation in our country (Shahriar, 2021).

As the customer has so many options for each product that they want to buy through F-commerce, the competition is very high. In addition, to sustain a business in this competitive F-commerce field, knowing customers' purchase choices, maintaining the availability of trendy or demanding products on one's Facebook business page is a crucial need of time.

An increasingly popular approach to rejuvenating and engaging younger, socially-conscious consumers is by utilizing social media. Social media components such as friends' gifting apps, purchase activity sharing with peers, and conversations are often incorporated. In the era of social commerce, users frequently transition from e-commerce websites to social networking platforms like Facebook and Twitter. Upon doing so, they often consent to

sharing some of their basic personal and demographic information, such as their location or Facebook likes, with the e-commerce business (Zhang & Pennacchiotti, 2013).

Hence predicting consumers' online purchases is of great importance. Therefore, this study will help to understand the taste of customers and the level of comfort ability they have while purchasing products online using an F-commerce site.

1.2 Present State of the Study

Predicting consumers' online purchasing taste and behaviour, especially for e-commerce sites has been a burning topic ever since people find their comfort in online shopping and such activity has become their lifestyle. So many e-commerce sites either have worked on predicting consumers' purchase behaviour or have shared their customer database with different academic students to present a model that helps the e-retailers to forecast the future purchase. In comparison to that, no such model has been yet developed to understand the purchase behaviour or to predict the future buying tendency of any F-commerce customer.

A significant study has been made to predict purchase behaviour based on social like (Zhang & Pennacchiotti, 2013), however, later a study also clarifies that with Facebook 'likes' and 'comments' an accurate customer segmentation based on gender cannot be feasible. Because a person of one gender deliberately exhibits characteristics of the opposite gender, for example, in an attempt to hide their personality.

Hence without directly focusing on individuals' Facebook profiles and monitoring their 'likes' and 'comments', in this project, F-commerce customers have been directly asked about their purchasing nature, to understand them in a better way. With such real-time data, forecasting future purchases and clustering customers based on their similarity would help us to draw a proper knowledge of customer segmentation and also would help to understand the comfortable level of the F-commerce consumer in case of buying products from Facebook pages.

1.3 Purpose of Predicting Purchase Behaviour

Like any other production industry or online business, knowing what might happen in sales in near future would always provide an extra cushion for the retailers. And satisfactory service that comes through the preparation of prediction, renders comfort ability to consumers.

Thus, this project focuses on forecasting future purchases by clustering customers. Based on the cluster segmentation, a classification of customer comfort level will also be shown.

Though clustering is the process of grouping things in such a way that those in the same group are more similar than the other groups, however; clustering customers in terms of their purchase taste is not an easy task due to the presence of more dissimilarity of individuals rather than similarity. The more the differences are there in the choice of each consumer the more scattered the real-time data gets. So to choose an appropriate clustering algorithm one should consider whether the algorithm scales to the dataset or not. Because not all clustering algorithms scale data productively, even if the machine learning dataset can contain millions of samples.

Hence, the specific aims of this thesis are as follows:

1. To identify the customers who share the same satisfactory level and purchase similar products and cluster them in a group
2. From the clustering, classify and predict their future purchasing tendency

1.4 Methodology

There are various ways available for clustering where each strategy is best suited for specific data distribution. In this study, distance-based centroid clustering is used.

Another unsupervised learning technique in clustering is- clustering in the R programming language. Where based on their similarity, the data set is partitioned into numerous groups termed clusters. Following data segmentation, several clusters of data are created. A cluster's objects all have similar properties.

R supports a wide range of clustering techniques. For instance, the R core (stats) includes a k-means function. Whereas the popular "cluster" package supports agglomerative clustering (Agnes), medoid-based clustering (pam), fuzzy clustering (fanny), and clustering of very large data sets (Clara). The "cluster" package also includes several cluster metrics, such as the silhouette, gap statistic, and dissimilarity measurements.

K-means, in which the cluster center is determined by the mean of the cluster member data. The efficiency of this algorithm is limited by its sensitivity to initial conditions and outliers.

Whereas, in **K-medoid** algorithms, such as PAM (Partition around Medoids), centers are defined by the actual data point which minimizes the cluster dissimilarity measures.

In this project, one of the most common clustering algorithms - **Partitioning around Medoids (PAM-** known as k-medoid**)** is used, which is a distance-based clustering approach. Though, in the past, the PAM algorithm suffered from inefficiencies on large data sets. However, in recent years various optimization techniques have been applied to make it a fast and effective clustering algorithm. Being a medoid-based PAM is more robust than k-means in handling data with outliers.

Real data was collected and analyzed by surveying original F-commerce users. Based on that survey data an idea was generated regarding 'preference products' for each user. People who preferred similar price-ranged products had been grouped by clustering algorithms K-medoid with Principle Component Analysis (PCA). For classification and prediction of the future behaviour, Support Vector Machine (SVM) was used.

1.5 Limitations of Existing Models

Predicting purchase behavior of consumers is not something new in this competitive era. There are already many models out there to predict customer's future move beforehand. But there always remain scope to work on an existing model and make it better by eliminating the limitations of that particular system.

The common constraints that the existing systems have are:

- Different product segments often get overlapped
- Flexibility or robustness is comparatively low
- Presence of a huge number of outlines in the dataset

Furthermore, there are other existing model for e-commerce consumer's purchasing behaviour. Also, this cannot be denied that Facebook has its own algorithm to target consumers but there also remains scope of further work.

The purpose of this project work, hence, is not to overcome the limitations of existing system only but to develop a model with real-time data and then predict the real customers to serve the real F-commerce owner.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

For online shopping decision-makers, purchase prediction plays a critical role in improving the customer experience, providing personalized recommendations, and increasing revenue. Many studies have looked into session log purchase prediction by analyzing user behaviour to forecast purchase intent after a session has finished.

Social media platforms like Facebook and Twitter have taken this purchase prediction to another challenging level. Business through Facebook is now a significant topic worldwide, so are the competitions among different Facebook business pages. Thus predicting purchase behavior is critical for a variety of user-centric applications in online retailing platforms, including personalized recommender systems (Chang, et al., 2017), user activity modelling (Zhou, Ding, Tang, & Yin, 2018) and resource management (Hu, Hall, & Attenberg, 2014). Purchase behavior based on 'clicked and non-clicked' or 'purchased and non-purchased behavior' (Jia, Li, Yu, & Wang, 2017) helps any e-commerce platform to grab their potential customers.

So far, functional magnetic resonance imaging (fMRI) with SVM (Wang, Chattaraman, Kim, & Deshpande, 2015), optimized sampling method with decision tree model (Yi, Wang, Hu, & Li, 2015), and frameworks based on dynamic pricing (Gupta & Pathak, 2014) are used in predicting f and m-commerce customers' purchasing behaviour.

Direct studies on the human brain with neuroimaging technique (Martino, et al., 2008), the role of hemispheric asymmetry on purchasing behaviour (Niklas, Outi, & Mikko, 2013), and multi-scale modelling of behaviour dynamics (Huang, et al., 2019) show us the procurement decision of the human mind. Studies on multi-product utility maximization (Zhao, Zhang, Zhang, & Friedman, 2017), using special features to predict from data (Zhao, Yao, & Zhang, 2016), and streaming recommender systems (Chang, et al., 2017), product classification (Yousef, Jung, Showe, & Showe, 2007), utilizing real-world data for predicting accurate results (Pan, Demiryurek, & Shahabi, 2012), and recurrent neural network-based model to

compress historical data (Mikolov, 2010) have drawn an accurate result with real data set in predicting purchasing behaviour.

Furthermore, deep neural networks (Wu, et al., 2016) and other machine- learning algorithms have been trained to predict purchasing behaviour based on Facebook likes as mentioned earlier.

A propensity to buy model analyses non-transaction customer data, such as how many times a customer DM (direct message) to a Facebook page or how the consumer interacts with one's Facebook business page, to determine which prospects are ready to make their first purchase.

Certain demographic data can also be factored into these models. They might compare gender, age, and zip code to other prospective purchasers in consumer marketing, for example. Relevant demographics in company marketing include industry, job title, and location.

However, with Facebook 'likes' and 'comments' an accurate customer segmentation based on gender cannot be feasible. Because a person of one gender deliberately exhibits characteristics of the opposite gender, for example, in an attempt to hide.

2.2 Benefits of Purchase Prediction

When one can anticipate how his customers will act, it provides him certain benefits like reducing customer churn (also known as customer turnover, or customer defection, or the loss of clients or customers), identifying and targeting high-value customers, meeting customer demand, and encouraging loyalty. Moreover, it helps to personalize the customer experience, reduce marketing campaign spending, get products quickly to markets and improve customer experience.

With all of these advantages, it is easy to see why the adoption of customer prediction software to inform organizations' customers and brand experiences has been rising. It is also critical for online retailers' experiences and contentment.

CHAPTER 3

THEORETICAL BACKGROUND

3.1 Introduction

The Internet does more than connecting two or more parts of the globe together. Especially when our life-style has become dramatically clock-based with every rushing second and availability of accessible internet is getting cheaper, mankind won't leave any stone unturned to use technology against time and make our daily routine more tech-based. From product to service purchase, every single activity can be done online nowadays. And social media apps like Facebook and Instagram are adding more value to this easiness of online purchase. Lately, it feels like doing business through Facebook pages has become a trendy fashion less and a necessity of time more! Facebook is not only connecting one Ecommerce site's customer to the other sites but also penetrating the whole market of F-commerce consumers. For cases like our country, rising percentage of unemployment, keenness of being self-dependent along with busy lifestyle and uneasiness of normal movement due to tremendous traffic, online shoppers especially F-commerce users as well as entrepreneurs are significantly rising. In fact, in recent days Bangladesh has over 300000 Facebook business pages or F-commerce platforms, which is surprisingly less than the Ecommerce platforms (Shahriar, 2021)

So in order to survive through such a competitive yet beneficial and most demanding virtual platform, one needs to play smart and learn the pattern of his consumer's purchasing behaviour.

3.2 Defining F-commerce

F-commerce means Facebook Commerce, which is an online business word that refers to the design and development of content and retail sites within the social networking platform Facebook. Simply put, it refers to the sale of goods and services on Facebook.com. So to say, it's e-commerce within the Facebook environment. It's a type of so-called "social commerce" that makes use of social media to sell a variety of things without requiring a huge investment. The Facebook store allows an individual to complete the purchase process in one of two

ways: completely within ones environment or starting on the social network and ending on the company's website.

F-commerce also refers to customers sharing information about purchases on Facebook, whether they use the Facebook "Share" feature on an e-commerce site or share information about purchases made at a physical point of sale. F-commerce is based on the idea of introducing a new type of online shopping: social shopping.

3.3 F-Commerce Customer's Buying Behaviour and Prediction

The process of projecting clients' forthcoming purchases is known as purchase prediction. It benefits enterprises in two ways in general. Internally, it enables them to efficiently and effectively manage demand and supply. As a result, stocking expenses and waste are reduced. On the customer side, it assists them in providing a superior customer experience. As a result, purchase prediction makes reliable forecasts using machine learning algorithms.

How an online customer purchases from any online platform can be predicted by monitoring a few patterns of purchasing behaviour of that very individual. Those behaviours can be acknowledged easily from social media platforms and can be used as tools to understand future buying possibilities beforehand. From liking and sharing to recommendation of any product to Facebook, or gathering data from browser cookies or user's historical activity, often visited website's notification, liking and disliking of product advertisement each and every action of F-commerce's users can be used as an evidence or powerful tool of knowing their future purchase.

But the probability of getting biased information using those tools is also a possibility that we should consider. There are also some people who don't want to show their actual online purchasing behaviour especially on Facebook! It could be due to maintaining a social status or showing off to colleagues and/or friend circles. Also, paid recommendation is pretty common nowadays. One can easily recommend a product which he generally won't prefer to purchase for himself, but the nature of his job may force him to do so!

Hence there always remain some gaps to actually predict customer's buying behaviour. So to get more effective and efficient user data, in this project a survey method is preferred to acknowledge customer's future purchase probabilities over customer's online activity.

3.4 Clustering

Clustering is the process of accumulating similar types of data or information together from a huge or large data set. Here similarity means either those data's patterns or functionality are same in nature than those of the other data present in that same data set. To put it another way, the goal is to separate groups with similar characteristics and keep them together in a group or cluster.

As characteristics and nature of humans vary from individual to individual, the pattern of their buying behaviour also varies. But still there are people who tend to buy similar types of product or service even if not the exact same product. For example, people who purchase clothes from Facebook also tend to purchase accessories or jewelleries. But maybe one's most purchased clothing is western where the other prefers to go with traditional. Again if someone feels comfortable buying plants from Facebook also feels comfortable buying books from Facebook; but they might not end up buying the same species of plants or same genre of books. So even if it is not possible to cluster that much accurately with varieties of similar products, it is possible to cluster those people together who tend to buy similar types of products. That, if not as a whole but at least to some extent helps to draw a pattern for customer buying behaviour.

In this project, with the help of a virtual survey among students and job holders, it is being tried to cluster those people together who tend to buy the same products. Later with another tool SVM accurate classification and prediction has been made on purchase behaviour.

3.4.1 Methods or Types of Clustering

There are two main categories of Clustering; Hard Clustering and Soft Clustering. But more or less there are 6 major methods to accomplish any of these two clustering. Typically, Hard Clustering involves assigning each data point to a single cluster, whereas Soft Clustering is more flexible. In Soft Clustering, the output or result is based on a predetermined set of cluster groups, allowing for some variability.

In cluster analysis, the initial set of objects is divided into groups based on similarity, and the groups are then labelled. Clustering methods roughly include **Partitioning-based,** and **Hierarchical-based** (Mondal & Choudhury, 2013). But there are also other clustering methods like, **Density-based, Grid-based, and Model-based** clusters.

Partitioning-based Clustering:

The process of partitioning items into k clusters involves creating distinct groups, where each cluster represents one partition. Certain conditions must be met, such as each cluster containing at least one data object and each data object being assigned to only one cluster. In partition-based clustering, all clusters are mutually exclusive. These strategies are designed to optimize a specific similarity function, with distance being the primary factor examined. The number of clusters (k) is typically defined by the user, and this type of clustering is usually unsupervised (Mondal & Choudhury, 2013).

The two most commonly used algorithms for this type of clustering are K-means clustering and k-medoid clustering or PAM (Partitioning around Medoids). Although both algorithms are designed for clustering, they have different internal structures. K-means clustering is primarily a distance-based clustering algorithm, while k-medoid or PAM operates differently.

K-means clustering:

It is one of the most extensively used and unsupervised method learning for unlabelled data i.e. data that cannot be defined as different categories or groups. Based on the distance metric used for clustering, it divides the data points into k clusters. The user has the freedom to choose a certain value for k. And then, distance between the data points and the cluster centroids is calculated. In this cluster process, k stands for the number of clusters calculated from the data set. Whatever number of groups the variable k represents from the data, the algorithm aims to find a match group for those data (Sharma, 2022).

The nearest data point of the centroid of a cluster gets the opportunity to tie itself with that cluster. It recounts the centroid of the cluster after each iteration. The process keeps going on up to a predefined number of iterations or, until the value of centroid of a cluster does not change after an iteration.

And this is the reason k-means clustering is often considered as the most difficult clustering process to implement in large data set. Because it keeps computing the distance between each data point and its centroid after a single iteration. In a cluster, each data point has a certain distance from the centre of the cluster. The purpose of the objective function of k-means clustering is to minimise that particular distance, recalculate the center and continue the process (Mondal & Choudhury, 2013).

K-medoid clustering:

On the other hand, there is another very popular algorithm for clustering known as k-medoid algorithm or PAM, i.e. Partitioning around Medoids, which have been used in our proposed model. It is a robust version of k-means and less sensitive to outliers.

The clustering process used in PAM is similar to k-means clustering, but with a different method for assigning the center of each cluster. In k-means clustering, the input data for each cluster does not have to be the average of all the data points, while in PAM, the medoid of the cluster is used as the input data. Hence the objective of PAM is to figure out the point that is closer to the center or that is the center itself of that cluster. In this process from the dataset of input, some are chosen arbitrarily and those chosen data are called 'selected' while the remaining is named 'non-selected'.

So it can be concluded that, the distance-based algorithm k-means tries to minimise the distance or the total squared error while in k- mediods, after selecting k numbers of data arbitrarily, a swap has been done between the selected and non-selected data point. To understand the swap effect, a cost is also related to the quality of partitioning the non-selected data to selected data. If the cost is less than 0, the selected data is to be marked as non-selected and the non-selected data to be marked as selected. And the process continues until the medoids keeps changing (Mondal & Choudhury, 2013).

CLARA (Clustering Large Applications):

CLARA is basically an augmented version of PAM that is good for large data set clustering because in this process the calculation time has been reduced for better effectiveness.

As a representative of actual data, this extended process of PAM targets a section of data irrationally from the entire data set. Then, by applying the PAM algorithm it multiplies the sample of the data. This process continues for quite a time. After a number of iterations it goes with the best cluster. Thus by choosing some unspecified data of the input data the CLARA process calculates the best medoids in the chosen sample rather than the whole dataset.

In order to deal with dense datasets CLARA works better than K-Medoids, because of the less time in computation. But it is also true that the efficiency of CLARA algorithms depend on the selective dataset and on their best K- medoids. If any representative dataset failed to find the best medoids CLARA algorithms won't work perfectly (Sharma, 2022).

CLARANS (Clustering Large Applications based upon Randomized Search):

And just because of the drawback the CLARA algorithm has, a new algorithm Clustering Large Applications based upon RANdomized Search is being instigated. Both of these two algorithms work similarly but the only difference is in their clustering process (Tyagi, 2021)

Hierarchical-based Clustering:

Just as the name suggests, in this clustering method every new cluster or children cluster is generated from the parent cluster or prior cluster and thus turns into a dendrogram. This method can be categorized into two more classes: *Agglomerative (bottom-up approach) and Divisive (top-down approach)* (Tyagi, 2021).

Agglomerative clustering:

The most popular clustering technique where each data point functions as an individual cluster at the very beginning and then a group is formed by each of the data points who are closest to each other (Tyagi, 2021).

Divisive clustering:

This is exactly the reverse of Agglomerative clustering. Accumulating all the data points this method at first completes the member for one cluster. Then divide that one particular cluster into more clusters. According to the linkage criteria, these algorithms conduct the linkage between the clusters and build distance matrix of all the current clusters. A tree-shaped diagram is used to show how the data points are grouped (Sharma, 2022).

Three types of linkages are explained; Single, Complete, and Average, each with a unique method for measuring the distance between two clusters.

Single Linkage: Single Linkage calculates distance based on the shortest distance between any two points in the two clusters.

Complete Linkage: Complete Linkage calculates distance based on the greatest distance between any two points in the two clusters.

Average Linkage: Average Linkage calculates distance based on the average distance between each point in one cluster and each point in the other cluster (Tyagi, 2021).

In addition, two examples of Hierarchical Clustering are provided below:

1. CURE (Clustering Using Representatives) and
2. BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies).

Density-based Clustering:

When similar types of data points are situated near each other and created dense population for a cluster, Density-based clustering acknowledges that destined region and alike data points. Not only that, this method also recognizes the different data points of less density region. And by the nature of a process this clustering method has a great ability to combine two clusters with good accuracy (Tyagi, 2021). The method has two examples:

- DBSCAN (Density-based Spatial Clustering of Applications with Noise)
- OPTICS (Ordering Points to Identify Clustering Structure)

These methods utilize distance measurements between objects to group them, resulting in spherical clusters that may make it difficult to identify non-spherical clusters. In addition, clusters are formed in all directions as long as the neighborhood density meets a certain threshold. Density-based approaches, on the other hand, examine a data point's entire density and consider factors that might impact it, making it more resilient to outliers and noise. Nevertheless, certain algorithms, such as OPTICS and DenStream, can automatically eliminate noise and detect clusters of any shape.

Grid-based Clustering:

As all the data points are shown in a grid-shaped structure, this method is called Grid-based clustering. The structure is formed with a finite number of data cells or grids.

The algorithms of Grid-based clustering method are more interested in the value space around the data points than they are in the actual data points. And that's why it takes a different general approach than the other algorithms. The complexity of computation is also lesser in this algorithm which is one of the biggest advantages. And this very characteristic makes this algorithm suitable for handling enormous data sets. Additionally, grid-based techniques perform proportionally to the size of the grid and take up far less space than a real data stream.

After dividing the data sets into cells, it computes the cell density, which aids in cluster identification (Tyagi, 2021). The following are a few algorithms that utilize grid-based clustering: –

Statistical Information Grid Approach (STING):

The data set is split into smaller and smaller subsets by STING in a hierarchical and recursive manner, with each subset containing a unique number of even smaller subsets. STING also captures statistical information from each of these subsets, which helps in promptly addressing requests. (Sharma, 2022).

Wave Cluster:

This technique utilizes wavelets to characterize the space where the data is present. The data space comprises of a signal with multiple dimensions that assist in identifying clusters. Clustering of data points is performed in areas where the signal has high amplitude and low frequency. The software recognizes these regions as clusters, while high frequency sections of the signal indicate the boundaries of these clusters. To detect densely populated regions in the transformed space, the original feature space can be modified through wavelet transformation (Sharma, 2022).

Clustering in Quest or **CLIQUE:**

CLIQUE is a clustering method that blends density-based and grid-based clustering techniques. It divides the data space into sub-spaces using the Apriori principle and then identifies the clusters by calculating the densities of the cells. It has the ability to discover any number of clusters in any number of dimensions and is not limited by a specific parameter for finding clusters of any shape. Compared to other methods such as K-means, DBSCAN, and Farthest First, it performs better in terms of execution speed, time, and accuracy (Sharma, 2022).

Model-based Clustering:

These techniques rely on probability distributions as the input data and utilize pre-existing mathematical models to fit and improve the data. They use conventional statistics to calculate the number of clusters (Sharma, 2022).

To ensure the clustering is robust, they incorporate noise and outliers when calculating standard statistics. There are two categories of clustering techniques based on how they form clusters: statistical and neural network approaches.

Examples of these techniques includes;

- MCLUST (Model-based Clustering)
- GMM (Gaussian Mixture Models)

Model-based algorithms that use statistical methods follow probability measures when determining clusters, while neural network algorithms link their input and output to unit-carrying weights.

3.5 Principal Component Analysis (PCA)

(Géron, 2017) states that **PCA** or Principal Component Analysis is the most common and famous method of dimensionality reduction process. R PAM algorithm uses this process for clustering. Selecting the proper hyper plane is the first criteria of this method before transferring training data into the low-dimensional hyper plane.

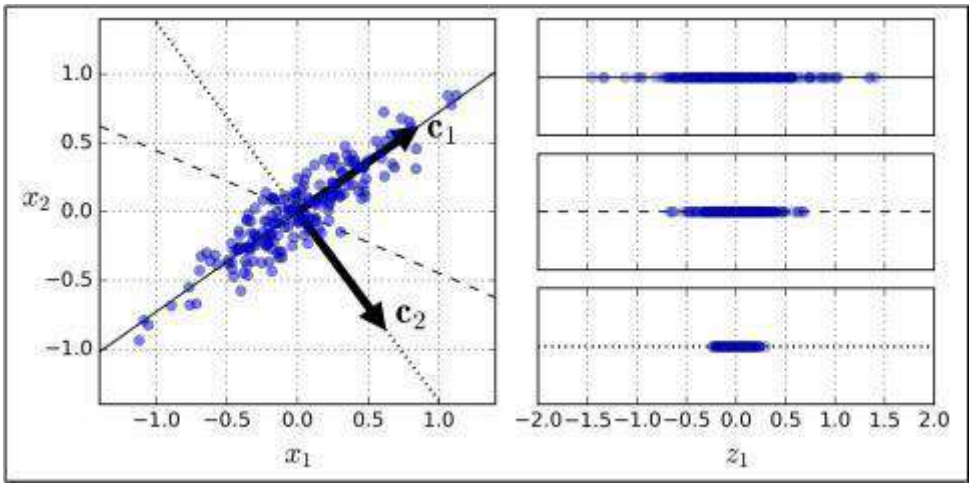

Figure 3.1: Choosing the subspace onto which to project the dataset

On the left side of Figure-3.1, a 2D dataset is shown with three different axes, while on the right side, the transfer of the dataset onto each axis is illustrated. The solid line projection preserves the highest variance, while the dotted line projection preserves the minimum variance, and the dashed line projection preserves a moderate amount of variance. It is logical to select the axis that preserves the most variance, as it would result in the least data loss. If the dataset had additional dimensions, PCA would discover new orthogonal axes. The axis that minimizes the mean squared distance between the original dataset and its projection onto that axis is another technique that supports this decision. This is the fundamental principle of PCA. However, the orientation of principal components is not stable, and slight changes to the training set may cause new principal components to point in the opposite direction of the original ones. Nevertheless, they usually remain on the same axes, and the plane defined by a pair of principal components may occasionally rotate or swap, although this is a rare occurrence (Géron, 2017).

One limitation of the PCA process is the instability of the principal component orientation. If there are any minor changes to the training set and the PCA process is rerun, some of the new principal components may point in the opposite direction from the original ones. However, they usually remain on the same axes. The plane defined by a pair of principal components may occasionally rotate or swap, but this is uncommon.

3.5.1 Different Uses of PCA (Géron, 2017)

PCA for Compression: PCA is an amazing tool that can help to compress data and reduce the original size of a given dataset. Also it is possible to decompress the data set again by applying inverse transformation of the PCA projection, though it will not get back to the original size of the dataset.

Incremental PCA: Incremental PCA (IPCA) helps to divide the training set into mini-batches and input each mini-batch to an IPCA separately. This is helpful for using PCA online and for big training sets. But there is a drawback with this process that the entire dataset does need to fit in the memory for the SVD (Singular Value Decomposition, standard matrix factorization technique for decomposing training dataset) algorithm to run.

Randomized PCA is a faster alternative for PCA, which is a stochastic method that quickly finds an approximation of the first d principal components. Scikit-Learn provides this option, which is much quicker than previous algorithms.

Kernel PCA is another option for PCA that uses the trick of mapping a complex nonlinear decision boundary in the original space to a linear decision boundary in the high-dimensional feature space. This allows for complex nonlinear projections for dimensionality reduction. This is called Kernel PCA (kPCA). After projection, it is effective at preserving clusters of instances or even unrolling datasets that are close to twisted manifolds, which is a useful feature of kPCA.

3.6 Support Vector Machine (SVM)

(Géron, Support Vector Machine, 2017) in this book it is also mentioned that SVM is crucial to understand the differences between classification and clustering. Using clustering, one can look for intriguing patterns in the data or confirm the purity of established classes. On the other hand classification is to newly discovered data records using supervised techniques and a database of previously identified data items. (Sarwar & Deb, 2022) , whereas mentioned that Support Vector Machine (SVM) is a well-known supervised machine learning technique that has been around for a long time and may still be utilized to solve classification problems requiring enormous quantities of data. It is particularly advantageous for multi domain applications that operate in a big data environment. It is, however, mainly employed to solve categorization difficulties. Using the SVM method, we can represent each data point in an n-

dimensional space, where n represents the number of features that a system has. Each feature's value is depicted as the value of a particular coordinate. Afterward, classification is performed by identifying the hyperplane that most effectively distinguishes between the two classes.

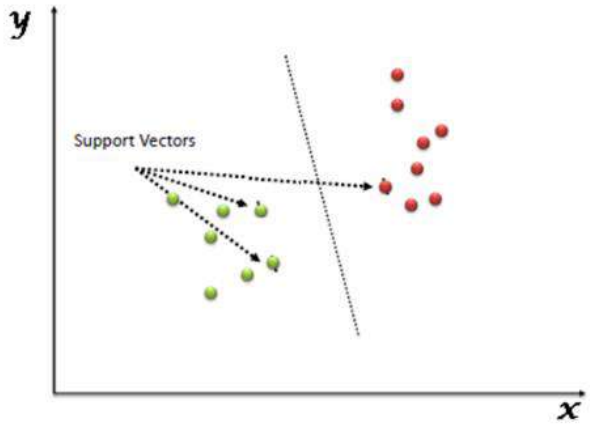

Figure 3.2: Support Vector Machine

The SVM classifier requires a training set of instance-label pairs (xi, yi), where i ranges from 1 to n. The input is labelled as xi ∈ Rn, and the output is labelled as yi ∈ {−1, +1}, the SVM classifier satisfies the following conditions:

$$\{w^T \varphi(x_i) + b \geq +1, \text{ if } y_i = +1\}, \text{ or } \{w^T \varphi(x_i) + b \leq -1, \text{ if } y_i = -1\}$$

W represents the weight of the vector, b represents the bias, and φ (•) represents a nonlinear function that maps the input space to a high-dimensional feature space. The linear separating hyperplane, denoted by $W^T\varphi(x) + b = 0$ distinguishes between two hyperplanes, with the margin width equal to (2/||w||2).

In classification problems where the data cannot be separated linearly, some misclassification is expected. To account for this, an additional variable ξi is included in the optimization model to allow for some misclassification. This variable is introduced into the primal optimization model as:

$$\underset{w,\, b,\, \xi}{\text{Min}} \quad \tfrac{1}{2} * w^T w + C \sum_{i=1}^{N} \xi i$$

Subject to,

$$\{y_i (w^T \varphi(x_i) + b) \geq 1 - \xi i, \; i = 1 \ldots N \; xi \geq 0, \; i = 1 \ldots N$$

Where C is a penalty parameter, a regularized constant that determines the trade-off between training error and model flatness.

The Lagrangian approach is utilized to tackle this quadratic optimization issue. To generate the Lagrangian function needed to determine the saddle point, Lagrangian multipliers I (i.e., support vectors) are introduced:

$$L_p(w, b, \alpha) = \tfrac{1}{2} * w^T w - \sum_{i=1}^{m} (\alpha_i y_i (w x_i + b) - 1)$$

By applying Karush Kuhn–Tucker (KKT) conditions for the optimum constrained function, Lp can be transformed to the dual Lagrangian LD (α):

$$L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \tfrac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i x_j)$$

The dual form of the primal optimization model can be transformed as:

$$\max_{\alpha} \; L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \tfrac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i x_j)$$

$$\text{Subject to} \quad 0 \le \alpha_i \le C \quad i = 1 \ldots m$$
$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

To transform the instance data into a high-dimensional feature space, the inner products in the objective function of the dual Lagrangian are substituted with the kernel function. It is crucial to select suitable kernel functions to ensure reliable classification results. The common kernel functions include linear, polynomial, radial basis functions (RBFs), and sigmoid. The optimal solution for the dual optimization problem is denoted by $\alpha*$. The decision function used for classification can then be defined as:

$$\text{Sgn}\left(\sum_{i=1}^{m} y_i \alpha_i * k(x, x_i) + b*\right)$$

The Kernel Function

The SVM algorithm employs a kernel function to perform a 2-dimensional classification of one-dimensional data. This kernel function is a mathematical technique that maps data from a low-dimensional space to a higher-dimensional space, as shown in Figure 3.3 below.

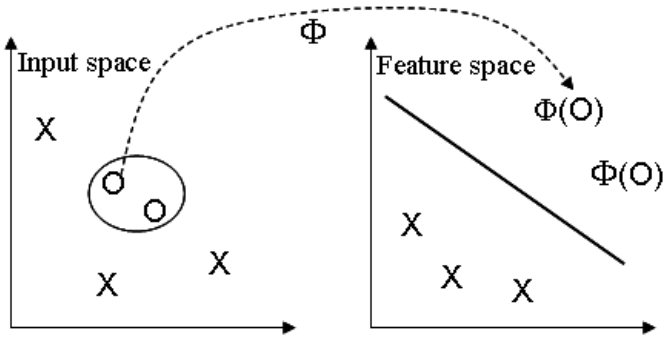

Figure 3.3**:** Kernel Function Mapping

$$\langle x1 \cdot x2 \rangle \leftarrow K(x1, x2) = \langle \Phi(x1) \cdot \Phi(x2) \rangle$$

Polynomial Kernel Function

The polynomial kernel function is a directional function, meaning that its output is determined by the direction of the two vectors in the lower-dimensional space. This behavior is attributed to the dot product used in the kernel. The magnitude of the output is also dependent on the magnitude of the vector $x_i K(x, x_i) = (1 + x \cdot XT_i)\, d$, 'd' is the degree of kernel function.

This versatile machine learning model can perform regression, outlier detection, and linear or nonlinear classification.

Despite the effectiveness of linear SVM classifiers in many situations, many datasets are not linearly separable. To handle such nonlinear datasets, one approach is to add polynomial features. While this is easy to implement and can work well with many machine learning algorithms, not just SVMs, it has limitations. At low polynomial degrees, it cannot handle very complex datasets, and at high polynomial degrees, it generates a large number of features that can slow down the model. However, using the polynomial kernel method can greatly improve the accuracy of classification. (Mondal & Choudhury, 2013).

CHAPTER 4

METHODOLOGY

4.1 Introduction

This chapter reflects the research methodology that is used to achieve the goals of this study. It showcases the data collection process by survey methods, data cleaning, clustering and finally classifying as well as predicting the results from real-time data of our customers. A distance-based clustering algorithm K-medoid is used for accumulating similar types of in a same group while Support Vector Machine algorithm has been used for predicting future purchase behaviour of the F-commerce users.

4.2 Procedure

The objective of this project was to first identify the people or F-commerce consumers who tend to purchase similar products and then predicting their future buying behaviour. We aimed to acknowledge the number of the active F-commerce customer and what type of products they were comfortable in buying from Facebook. It could have been much easier to predict data or to draw an estimated number of customers who purchase from online, especially from E- commerce sites, but to be more accurate in the data collection process an online survey has been chosen.

The following steps were followed thoroughly while analysing the dataset:

1. A survey has been done virtually and 139 participants responded by Google form questionnaires
2. The real time data of that survey was then filtered again. And in the process of data cleaning, 128 participants were found who respond to all the questions of the survey. People who did not answer maximum survey questions or who were not F-commerce customers were removed from the list.
3. And all the nominal data were then converted into numerical value by using ordinal integer encoding. A user preference product idea was developed based on the survey data after converting the nominal data to numerical data
4. There were four types of products with a lower and higher value range of each. From the result of the survey a pattern has been discovered that, a person who buys books in

a low range from Facebook also tends to buy clothes or gadgets at the lowest possible price. And if someone prefers high ranged products, he is more likely to buy every category of the product at a high price.

5. Later on, all of those people who tend to buy similar price-ranged products are clustered together by using an unsupervised **Partitioning around Medoids (**PAM) algorithm. Which is a distance-based clustering algorithm that works as K- medoid with Principle Component Analysis (PCA) process.
6. Eventually, the machine was trained with a number of trained data to understand which customers are pleased to purchase from Facebook and which customers aren't. Pleased customers were weighted with value 1 and unsatisfied customers were weighted with value 0. Later on from the tested data, a supervised machine learning algorithm Support Vector Machine (SVM) then predicts the future purchase-satisfaction of the customers.

4.3 Clustering

Technical Steps for Clustering:

- Data Pre-processing
- Elimination of missing values
- Converting nominal values to numerical values by Integer ordinal encoding
- Standardisation of variables
- Data scaling
- Apply PCA (Principal Component Analysis)
- Selection of optimum clusters using elbow method, gap statistics and silhouette
- Plotting the clusters
- Silhouette width comparison and checking
- Final decision

As depending on a single method could mislead the final outcome, while selecting the optimum number of clusters, three methods- i) Elbow, ii) Gap statistics, and iii) Silhouette, were used explicitly to verify the credibility of the solution.

4.4 The Libraries Used in Clustering

The libraries that have been used throughout using R-Studio are below:

i) caret, ii) ggplot2, iii) cluster, iv) factoextra, and v) alluvial

Acquire data

The actual data that has been collected from Facebook survey has a sample size of 128. That is 128 people among the all who responded to the survey were surely Facebook users and most of them were F- commerce customers. Among those 128 data from the dataset some data were first trained and the rest were then tested as per the instruction of training data.

Train data set size = 90 and Test data set size = 38

4. 5 Classification and Prediction

Data Background

From the 139 responses of the Google survey form 128 people were F-commerce consumer. Hence it can be said that a total of 128 instances were actually included in the survey dataset. Those instances provide information about how F-commerce users prefer different genres of products and how their preferences change as the products' prices vary.

In the survey, there were 8 categories of products (Fig. 4.1). But data of some product categories were way too much scattered and small in number that all products were categorized into four categories (e.g. Books, Plants (pot plans), Gadgets, and Apparel or Clothing).

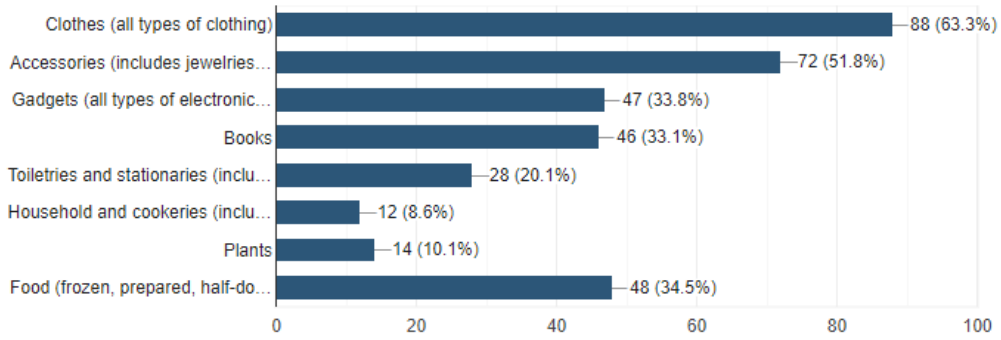

Figure 4.1: Categories of products that were present in the survey questionnaire

Using ordinal integer encoding, categorical variables were converted to numerical data. For example, a customer who prefers to buy a book within BDT 300 is given less weight than someone who prefers to buy a book with a higher price.

In the output section, Customers with a higher probability of buying products from F-commerce sites are assigned a binary classification value of 1 while others are assigned a binary classification value of 0. A customer with a satisfaction level of 1 is considered highly satisfied with their online shopping experience. These customers are likely to buy more products from F-commerce sites. This conclusion is based on their buying frequency and relative preference to prices of products.

CHAPTER 5

SURVEY AND DATA REPRESENTATION

5.1 Introduction

Online purchase has been a regular practice in Bangladesh over the year. However, in recent days, because of our livelihood and lifestyle our shopping pattern has gone through a tremendous shifting. Moreover, Facebook business added a completely new dimension to the online purchase process.

Where there are people who feel comfortable to buy their necessary things from online, there are also another group who do not find the virtual shopping platform (especially F-commerce platform) as a reliable one.

Interesting part is the people who prefer to purchase from Facebook can be categorised into another two categories. One are those whose purchase decision depend on view of Facebook live or content reaction that is biased, and the other are those who genuinely buy the product from the feel of their need.

As human mind and decision making procedure is indeed a complex process, a real survey has been done through online to collect as much clean data as possible.

During the survey the participants were strongly instructed to not to consider any E-commerce site and page (i.e Daraz, Chaldal, Rokomari, Ajkerdeal, Homential, Aliexpress etc.) as Facebook business page.

5.2 Survey Questionnaire

The survey questionnaire was divided into two section. One for the Facebook product or service consumers (F-commerce customers) and the other one for the Facebook product or service provider (owner of F-commerce business page).

5.2.1 Questionnaire for F-commerce user or customers

The following questions were prepared for the F-commerce user or consumer:

1. Please write down your full name below.
2. Your age limit

i. 18-25
ii. 26-35
iii. 36-45
iv. 46-55
v. 55+

3. What is your current profession? (I.e. teacher, lawyer, student, writer etc.)
4. For online purchase, how frequently do you shop from Facebook based business pages?
    i. Very frequently
    ii. Frequently
    iii. Only occasionally
    iv. In case of emergency
    v. Do purchase online frequently but never from Facebook pages
5. What type of products do you generally look for/buy from Facebook based business page? (you can choose more than one option)
    i. Clothes (all types of clothing)
    ii. Accessories (includes jewelleries or ornaments, shoes, bags, watches, bands or clips etc.)
    iii. Gadgets (all types of electronic or mechanical products/tool)
    iv. Books
    v. Toiletries and stationaries (includes soap, shampoo, floor cleaner, scale, pencil, pen etc.)
    vi. Household and cookeries (includes bed, sofa, almirah, pan, catering set, cabinet etc.)
    vii. Plants
    viii. Food (frozen, prepared, half-done, dried, canned, animal food etc.)
6. For clothing and accessories you look for product brand-
    i. Always
    ii. Mostly
    iii. Often
    iv. Not really brand oriented shopper
    v. Never
7. What is your in general preferred price range while purchasing clothing and accessories from Facebook based business pages?

i. Not up to 3k
ii. 4k-8k
iii. 10k-15k
iv. 16k-25k
v. In case of pre-order more than 25k
vi. Don’t buy such items from Facebook pages

8. How comfortable are you to buy clothes and accessories from Facebook pages?
    i. Extremely comfortable
    ii. Pretty comfortable
    iii. Quite comfortable
    iv. Moderately comfortable
    v. Not at all comfortable
9. What is the preferable price range for buying gadgets from Facebook based business pages?
    i. 3k-5k
    ii. Up to 10k
    iii. Up to 25k
    iv. Up to 30k
    v. In case of pre-order more than 50k
    vi. Don’t buy gadgets from Facebook pages
10. How preferable do you think buying gadget from Facebook pages is?
    i. Extremely preferable
    ii. Pretty preferable
    iii. Quite preferable
    iv. Moderately preferable
    v. Not at all preferable
11. What is your preferable price range for purchasing books from Facebook based business pages?
    i. 200-500 taka
    ii. Up to 1000 taka
    iii. 2000- 3000 taka
    iv. 3100-5000 taka
    v. In case of pre-order purchased, more than 5000 taka
    vi. Don’t buy books from Facebook pages

12. If you have to rate your comfort level of buying books from Facebook pages, it would be-
    i. Extremely comfortable
    ii. Very comfortable
    iii. Quite comfortable
    iv. Not that much comfortable
    v. Not at all comfortable
13. While purchasing plants from Facebook based business pages your preferable budget-
    i. 50-350 taka
    ii. Up to 500 taka
    iii. 600- 1000 taka
    iv. Up to 2000 taka
    v. 2500-3500 taka
    vi. Don't buy plants from Facebook pages
14. Do you like to buy plants from Facebook based business page?
    i. Absolutely
    ii. Pretty much
    iii. Moderately
    iv. Not that much
    v. Do not like at all
15. Review your experience of purchasing quality goods/products from Facebook based business page-
    i. Always get the promised quality products
    ii. Mostly get promised quality products
    iii. Generally get desired quality products
    iv. Very often get desired quality products
    v. Never satisfied with the quality of purchased product
16. What payment method do you follow in general while buying from Facebook business page?
    i. Online payment using cards
    ii. Online payment using apps (i.e. Bkash, Nagad, rocket etc.)
    iii. Cash in delivery
17. In case of online payment, how much comfortable are you with your payment method?

i. Very much
ii. Quite comfortable
iii. Moderately comfortable
iv. Always in tension of being cheated but do it anyway
v. Not at all comfortable because of becoming a victim of fraud several times

5.2.2 Questionnaire for F-commerce owner

The following questions were prepared for the F-commerce owner:

1. If you are a Facebook business page owner, how many anniversaries have your page celebrated?
2. What type of products do you promote through this business? / What type of products do you sell?
3. Who are your regular customers?
    i. Women
    ii. Men
    iii. Girls
    iv. Boys
    v. Both male and female
4. How many people have liked/reached your page till date? (i.e. 5k, 1M etc.)
5. How satisfied are you with your customer?
    i. Extremely satisfied
    ii. Very satisfied
    iii. Satisfied
    iv. Moderately satisfied
    v. Not at all satisfied
6. Your page have created a fan-base for-
    i. Aged people (45+ years old)
    ii. Middle aged people (35-45 years old)
    iii. Adult (20-34 years old)
7. What are your preferable payment methods?
    i. Online payment with cards
    ii. Online payment with apps (i.e. Bkash, Nagad, rocket etc.)

iii. Cash in delivery

8. In case of online transaction, have you faced any trouble from customers?
    i. Never
    ii. Very rare
    iii. Sometimes
    iv. Often
    v. Always

5.3 Data Representation

5.3.1 Data representation for F-commerce user or customers

a) Data representation of the survey participants in terms of age

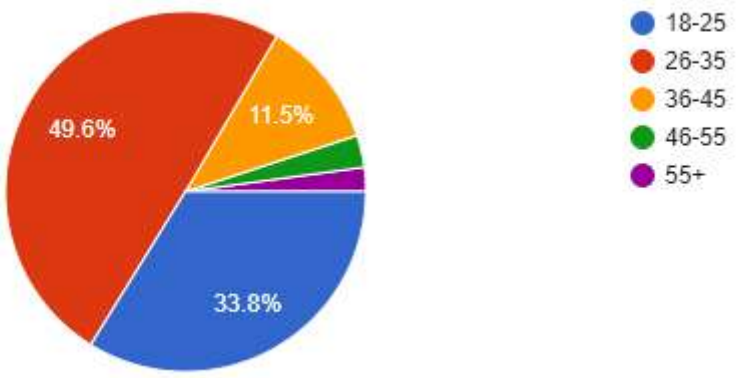

Figure 5.1: Age range of the survey participants

Among the 139 participants, 69 people (49.6%) were of the age group of 18 to 25. Whereas 47 (33.8%) people claim to be within age range 26-35. A few people (16.6%) were from the other age groups.

It is better to mention here that, in the data cleaning process, there were only 128 actual data; that means 128 people either answer all the questions perfectly or were a true consumer of F-commerce. Hence the percentage shown here is not completely correct in terms of sample size. Because it is based on the sample size of 139, whereas in reality, a sample size of 128 was actually taken.

b) Data representation of the survey participants in terms occupation

From figure 5.2 it is clear that the occupation part was an open question in the survey form. Hence people were free to mention the exact designation (i.e. businessman, freelancer, army officer etc.) of their occupation.

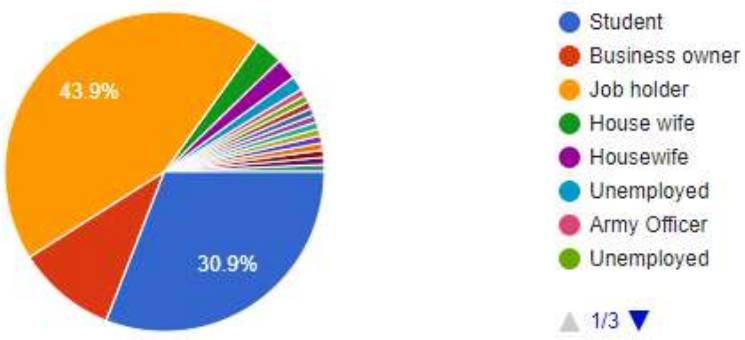

Figure 5.2: Different occupation of the survey participants

However, for the ease of the data classification, the entire participants were grouped into "Student "and "Job Holder" these two categories only. It was predicted that, an army officer, freelancer, businessman or private job holder all were grouped under the same occupation category i.e. "Job Holder".

c) Data representation of the survey participants in terms purchasing frequency

Figure 5.3: F-commerce consumer's buying frequency

As mentioned earlier, this very figure also gives us a clear indication that not every participant was an F-commerce consumer. However from the very frequent user to the occasional use there were 128 people as F-commerce customers.

d) Data representation of survey participants' preference in terms of F-commerce product category

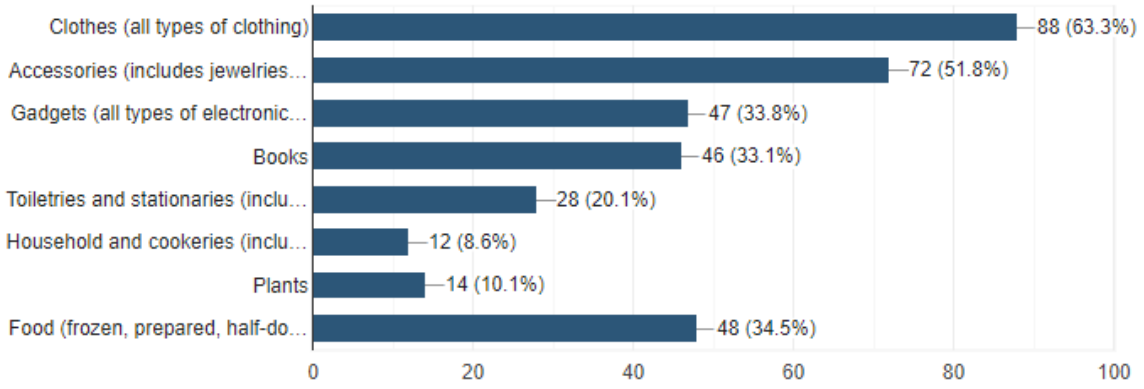

Figure 5.4: Preference of purchasing different F-commerce products

In chapter 4, figure 4.1 states that, even if we have around 8 categories, maximum F-commerce users show their interest in only four categories. The other categories, like household and cookeries, food, toiletries and stationaries were small as data and very scattered in nature. And even if 72 people claimed to be the consumers of accessories, unfortunately most of them did not properly complete the survey. Those incomplete data would left this model with so many empty cells in the raw result sheet (excel sheet) or missing values. That might have led to ambiguous result. Hence, those data were omitted during data cleaning process.

e) Data representation of survey participants in terms of quality satisfaction

Figure 5.5: Level of satisfaction of F-commerce consumer

From figure 5.5 it can be stated that, 83.4% people claimed to be either highly satisfied because of the good quality of products or services they get from F-commerce based business page. And the rest 16.6% were not satisfied with the quality.

5.3.2 Data representation of buying preference of students and job holders

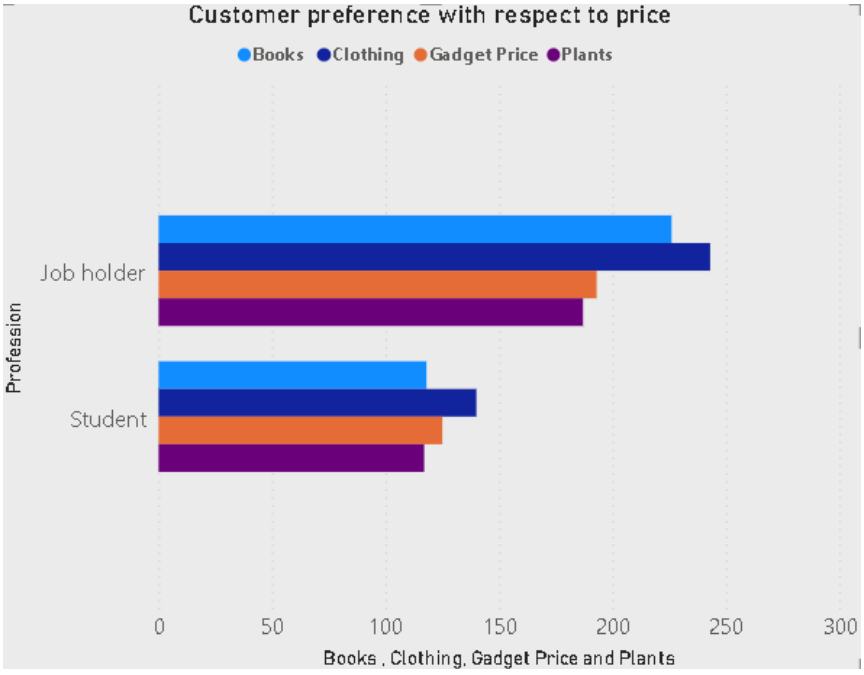

Figure 5.6: Buying preference of customer with respect to price

From figure 5.6 it is clearly visible that, in all the four categories, job holder prefers to purchase more of these products than students. And this is very much logical in terms of pricing. Another interesting fact can be concluded from above shown picture that is both students and job holders tend to buy clothes more than any other categories. Figure 5.4 has already given a clear picture of this statement before, it graphically shows that about 63.3% F-commerce consumers prefer apparel over any other categories.

5.3.3 Data representation for F-commerce owner

Figure 5.6 states that there were very few people from the participants who had a Facebook business page. Even if their page follower rate or live views were higher and they have completed up to 3 years of their online business, but the percentage of such owners were so less (up to 5.4%).

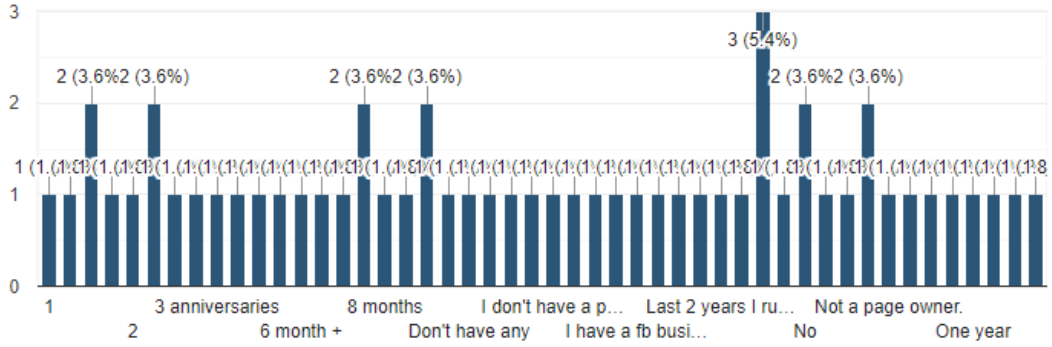

Figure 5.7: Active years of F-commerce business pages of the owner

Hence conversion of those nominal value into numeric number and then work with those data was not a practical decision in terms of building a machine learning model. For this reason, that entire segment (data cleaning of F-commerce owners) was out of consideration in this model.

5.4 Conversion of nominal data to numeric value

Ordinal integer encoding method was used throughout the process of data conversion from nominal to numerical. Because from age to price range the relation between all the data were ordinal, that is they could be categorized from smallest to largest.

a) Age range conversion

Age	
18-28	1
29-55	2
55+	3

Table: 5.1: Conversion of nominal data to numeric value in terms of age

From the above table, it can be stated that the people who belong to age range of 18 to 28 were given a weightage value of 1, while 29-55 aged and 55+ years old were provided with a value of 2 and 3 respectively.

b) Clothe price range conversion

Clothe Price Range	
Less than 1000	0
1001 to 1500	1
1501 to 2000	2
2001 to 3k	3
3001 to 4k	4
4001 to 5k	5

Table: 5.2: Conversion of nominal data to numeric value in terms of clothe pricing

Table 5.2 shows that clothe prices that were in between 1000 were scored as 0. The range from 1001 up to 1500 and 1501 to 2000 were weighted as 1 and 2 respectively. Whereas, price range between 2001 to 3000 and 3001 up to 4000 were scored as 3 and 4 respectively. Finally a weightage of 5 was given to those clothes that were up to 5000 BDT.

c) Book's price range conversion

Books' Price Range	
Less than 250	0
251 to 400	1
401 to 550	2
551 to 700	3
701 to 850	4
851 to 1000	5

Table: 5.3: Conversion of nominal data to numeric value in terms of book pricing

Table 5.3 provides a clear view on the weightage of book's pricing range. The books that were less than 250 were scored as 0. The range from 251 up to 400 and 401 to 550 were weighted as 2 and 3 respectively. Whereas, price range between 551 to 700 and 701 up to 850 was scored as 3 and 4 respectively. Finally a weightage of 5 was given to those clothes that were from 851 to 1000 BDT.

d) Gadget's price range conversion

Gadget's Price Range	
Less than 1000	0
1001 to 1500	1
1501 to 2000	2
2001 to 3k	3
3001 to 4k	4
4001 to 5k	5

Table: 5.4: Conversion of nominal data to numeric value in terms of gadget pricing

From table 5.4 it is clear that gadgets that cost in between 1000 were scored as 0. The range from 1001 up to 1500 and 1501 to 2000 were weighted as 1 and 2 respectively. Whereas, price range between 2001 to 3000 and 3001 up to 4000 were weighted as 3 and 4 respectively. Finally clothes that were up to 5000 BDT was given a weightage of 5.

e) Plant's price range conversion

Plants' Price Range	
Less than 90	0
91 to 150	1
151 to 200	2
201 to 300	3
301 to 450	4
451 to 500	5

Table: 5.5: Conversion of nominal data to numeric value in terms of plant pricing

Table 5.5 states that, plants that were less than 90 were weighted as 0. The range from 91 up to 150 and 151 to 200 were weighted as 1 and 2 respectively. Whereas, price range between 201 to 300 and 301 up to 450 were respectively weighted as 3 and 4. Finally a weightage of 5 was given to those clothes that were from 451 to 500 BDT.

CHAPTER 6

RESULT AND DISCUSSION

6.1 Result and Discussion

6.1.1 PCA and Clustering (Unsupervised Learning):

Step 1: Statistical Analysis:

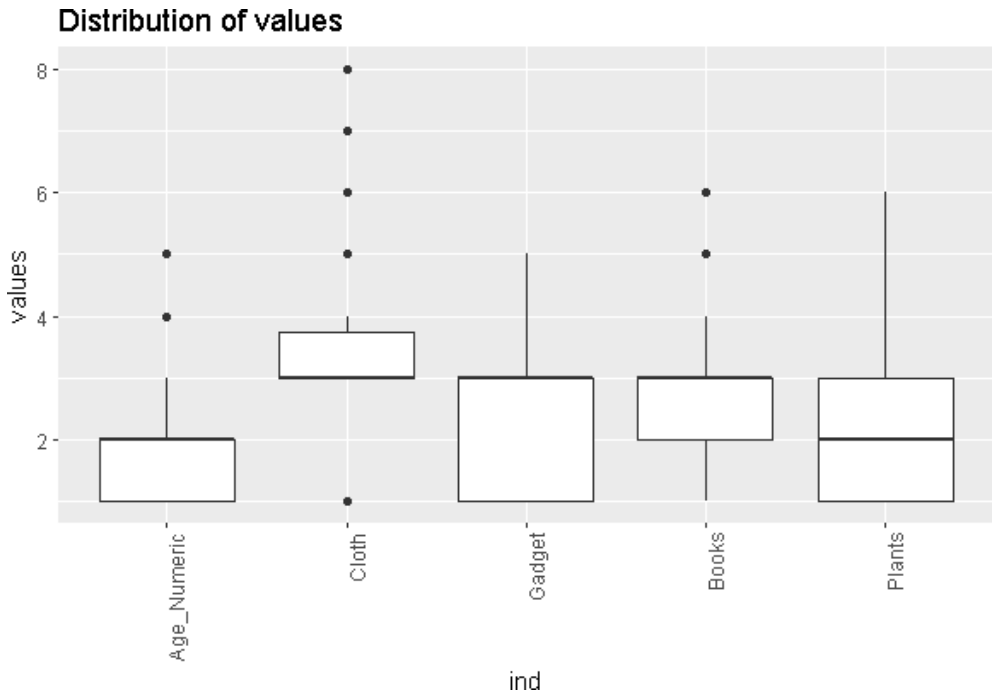

Figure 6.1: Distribution of all values in Box Plot

In the first step, box plots had been of all raw values of the dataset. Then, proper scaling was done to handle highly varying magnitudes or values. From the above figure 6.1, it can be seen that the variables gadgets, age, and plants have a wide range of values. To prevent an ML algorithm from giving more weight to larger values and less weight to smaller values, regardless of the unit of measurement, scaling was necessary. Additionally, feature scaling helps to normalize the range of features in a dataset. Since the survey data used in this study is a real-world dataset that is scattered and noisy, feature scaling was applied to standardize the data. In general, a certain type of standardization is preferred over normalization.

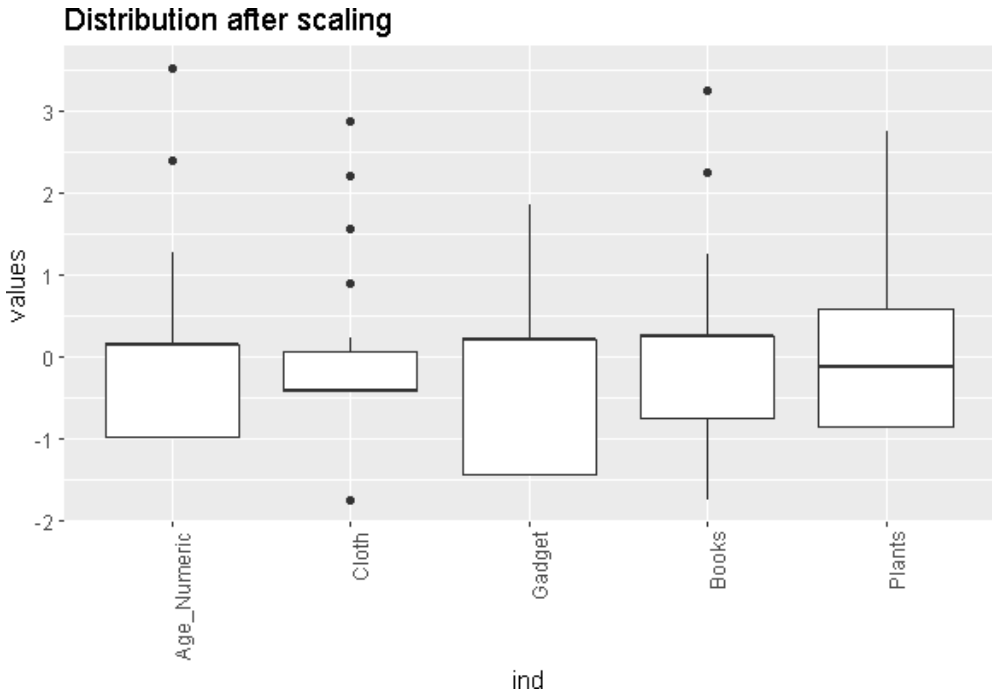

Figure 6.2: Distribution of all values in Box Plot after scaling

Majority of the machine learning model converts the data from input to output. Also, data point distributions are not always the same for different attributes of any dataset. Larger discrepancies between the data points of the input variables can lead to ambiguous results. That's why it is important to use scaling. The data points that are far apart from each other, scaling draws them closer. Hence it can be said that scaling is a method of securing accurate results of a model by placing all the data with a minimum gap in between them. So, in statistics or machine learning model approach data scaling is an inevitable method while working with large amounts of data.

Before performing PCA, the features in the dataset is scaled because PCA is sensitive to the scale of the features. The Standard Scaler from Scikit Learn has been used to standardize the features to have a mean of 0 and a standard deviation of 1, which is a common requirement for optimal performance of many Machine Learning algorithms. We validated our results by doing this step.

Step 2: Analysis of Variance:

```
> robpca.model <- princomp(sel.p,cor = FALSE,covmat = MASS
> cat('Loadings of the selected PCs: ')
Loadings of the selected PCs:
> prmatrix(round(robpca.model$loadings[,1:5],5))
             Comp.1   Comp.2   Comp.3   Comp.4   Comp.5
Age_Numeric 0.03257  0.23206  0.11123  0.93044  0.25885
Cloth       0.53642 -0.40309  0.35163  0.21153 -0.61756
Gadget      0.47112 -0.42741 -0.69690  0.08432  0.32028
Books       0.43194 -0.03698  0.57908 -0.25377  0.64212
Plants      0.55014  0.77435 -0.20731 -0.13430 -0.19160
> cat("\nNumber of PCs: ",ncol(robpca.model$loadings))

Number of PCs:  5
```

Figure 6.3: Principal components analysis in R Studio

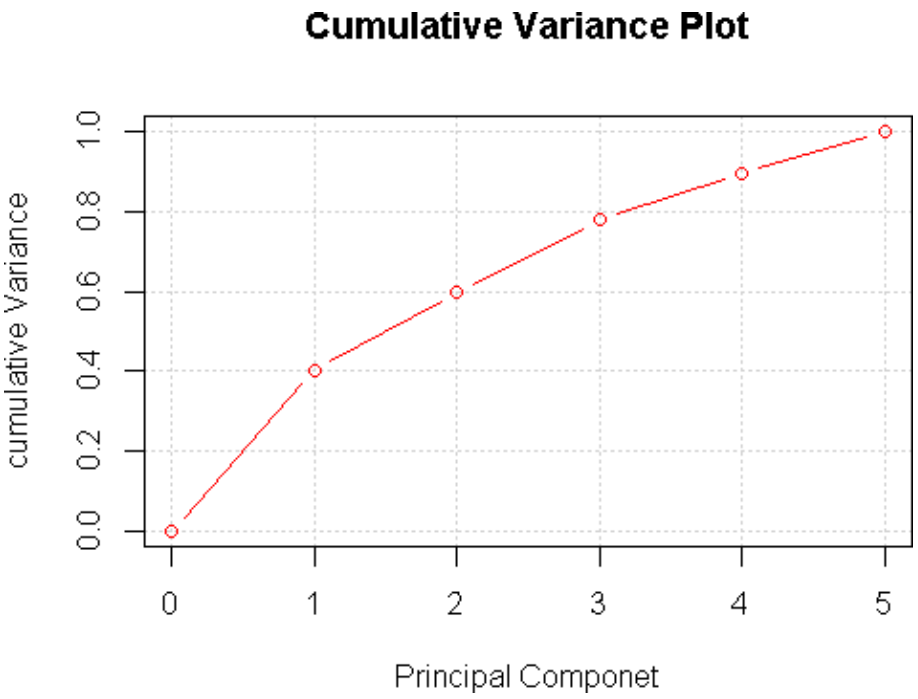

Figure 6.4: Cumulative variance plot of all PCs

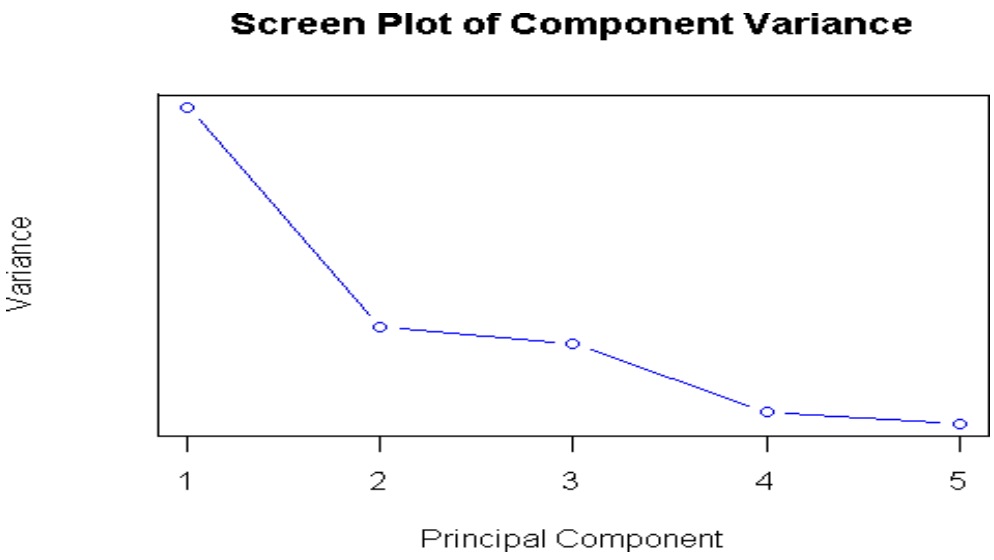

Figure 6.5: Screen Plot of component variance

From the above figures, we can observe that Principal components 1, 2, and 3 are responsible for 80% of variance in the dataset. Similarly, there is not so much variance in the no. 5 variable.

Step 3: Calculation of Silhouette width:

Silhouette width is a Machine Learning technique that is used to evaluate the consistency of clusters in data. The range of the silhouette width is between -1 and +1. A higher silhouette width score indicates that the object is well-suited to its cluster but not to its neighboring clusters. In other words, it helps in interpreting and validating the accuracy of the clustering algorithm.

```
data.frame(clus_no=1:kNo, cluster.model$clusinfo, avg_sil_width=cluster.model$silinfo$clus.avg.widths)
clus_no size max_diss  av_diss diameter separation avg_sil_width
      1   48 3.612861 1.693388 6.073024  0.9993141     0.1647161
      2   20 2.361528 1.521700 3.949890  0.9993141     0.2872406
      3   22 3.821388 1.989066 5.878853  1.3181172     0.1499952
# Cluster statistics for individual data points
sil.widths <- cluster.model$silinfo$widths
head(sil.widths[order(as.numeric(row.names(sil.widths))),], 3)
cluster neighbor   sil_width
      1        2  0.14328324
      2        1  0.42479977
      1        2 -0.03444636
print(paste("Sizes: silhouettes = ", length(cluster.model$silinfo$widths),
            "; clustered set = ", length(cluster.model$clustering),
            "; original data set = ", nrow(sel.p)))
1] "Sizes: silhouettes =  270 ; clustered set =  90 ; original data set =  90"
print(paste("Average cluster silhouette width: ", round(cluster.model$silinfo$avg.width, 2)))
1] "Average cluster silhouette width:  0.19"
```

Figure 6.6: Silhouette width calculation with train dataset

From picture 6.6 it can be concluded that, the average silhouette width of the dataset is 0.19. For cluster 1, the average silhouette width is 0.1647161, for cluster 2 the value is 0.2872406, and finally for cluster 3, the value came out as 0.1499952. By doing average of the three, we have got a value of average silhouette width: 0.19.

```
kNo <- 3
std <- FALSE
# Create a cluster model
cluster.model <- pam(x=trans.p, k=kNo, stand = std)
data.frame(clus_no=1:kNo, cluster.model$clusinfo, avg_sil_width=cluster.model$silinfo$clus.avg.widths)
clus_no size max_diss  av_diss diameter separation avg_sil_width
      1   48 3.612861 1.693388 6.073024  0.9993141     0.1647161
      2   20 2.361528 1.521700 3.949890  0.9993141     0.2872406
      3   22 3.821388 1.989066 5.878853  1.3181172     0.1499952
```

Figure 6.7: Details of clusters

Step 4: Calculation of number of clusters:

Calculation of Number of clusters with three different methods are shown below:

Silhouette Method

The silhouette method is a technique used to determine the ideal number of clusters in a dataset. It assesses and explains the consistency within the data clusters. By calculating the silhouette coefficients for each point, it determines how closely a data point belongs to its own cluster compared to other clusters.

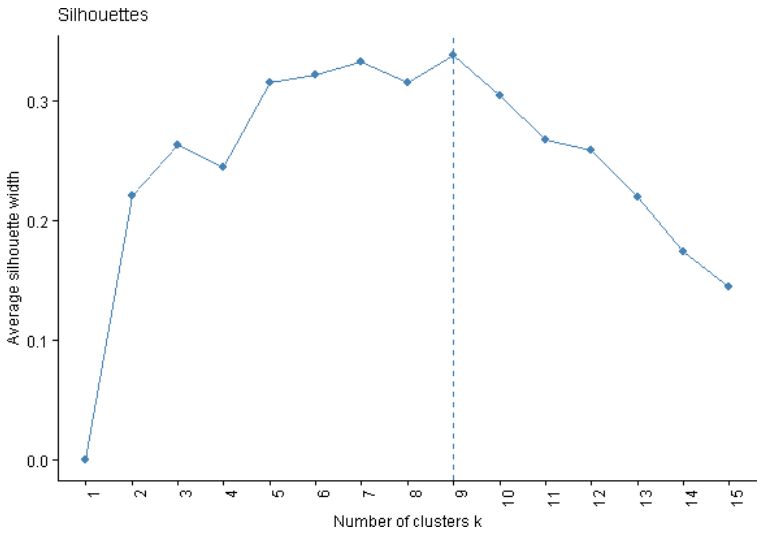

Figure 6.8: Silhouette method

From the x-axis of figure 6.8-graph it is clearly visible that just after the point 3 the curve goes down suddenly and then moves up. Before this sudden fall, the data of the graph went up without showing much dissimilarity in the path pattern.

Gap Statistics

Gap statistic is another effective method for determining the most possible number of clusters in a partition clustering.

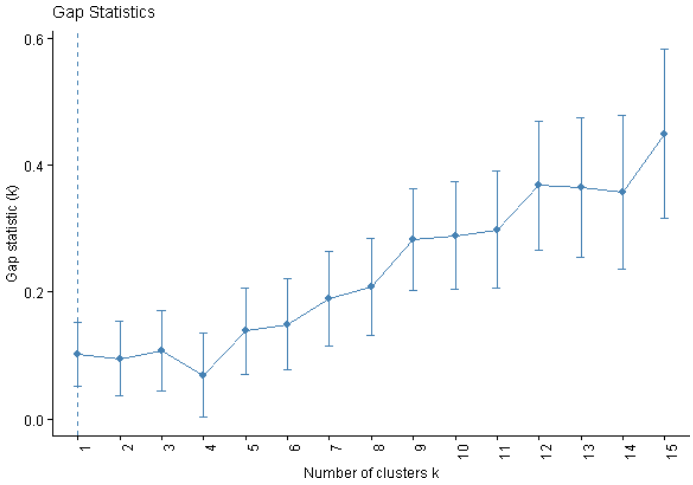

Figure 6.9: Gap Statistics methodology

Just like the Silhouette method, Gap statistics method also shows the sudden fall of the graph right after point-3 on x-axis.

WSS Elbow Method:

This approach is widely used to determine the ideal number of clusters, and it relies on computing the Within-Cluster-Sum of Squared Errors (WSS) across varying cluster quantities (k), ultimately selecting the k where the reduction in WSS begins to plateau.

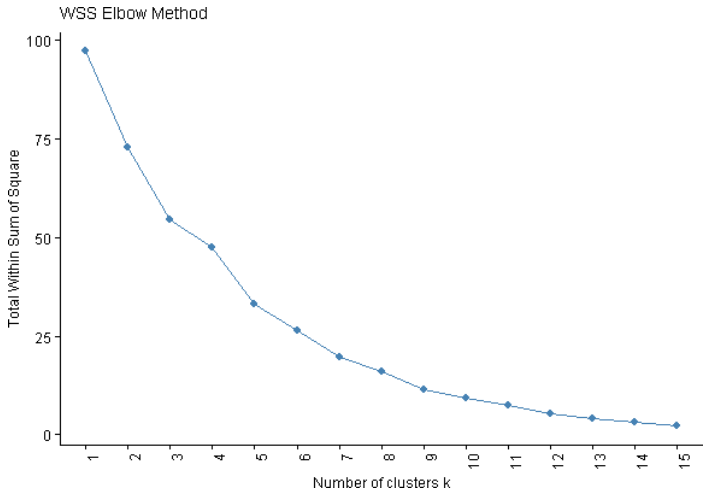

Fig 6.10: WSS Elbow Method

In figure 6.10, again just after point 3 along x-axis the curve bend like an elbow at the point 4. And the created a path smoothly downwards.

From the three above-mentioned methods, the optimum number of clusters had been found. For a proper validation 3 different clustering methods had been used to identify the numbers of cluster.

Step 5: Train Data Clustering:

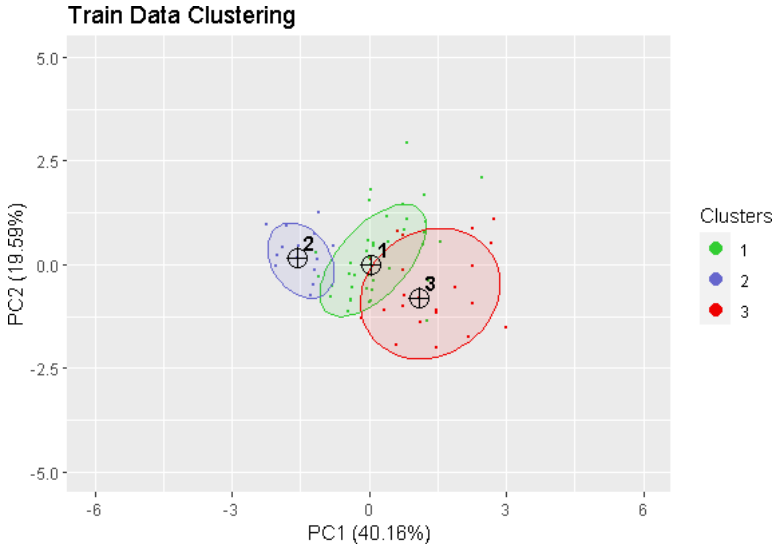

Figure 6.11: Train data clustering-1

More overlaps are being noticed between the cluster 1 and cluster 3, rather than cluster 2. That may happen due to the similar types of data between the clusters and because of less of variance, which may degrade the quality of clusters and may lead to error in real time marketing strategy in F-commerce business practices.

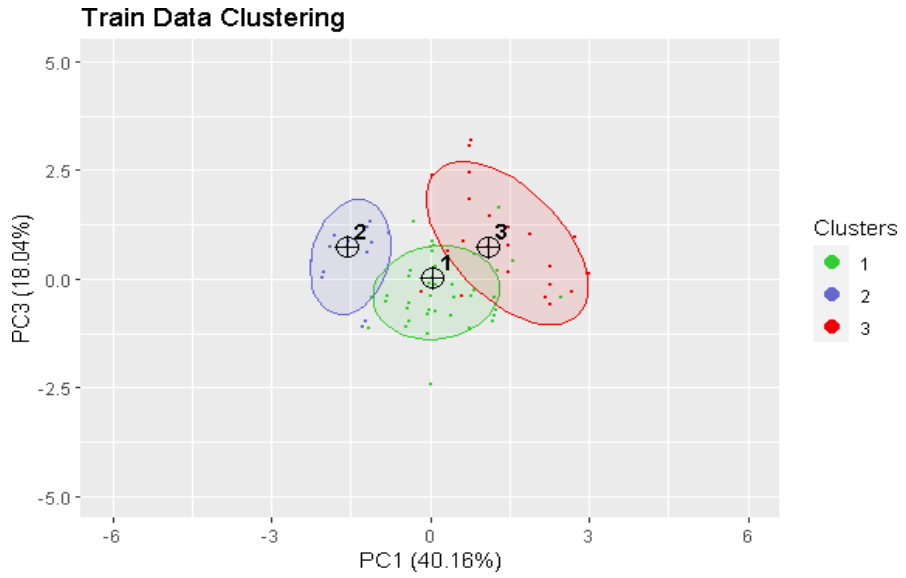

Figure 6.12: Train data clustering-2

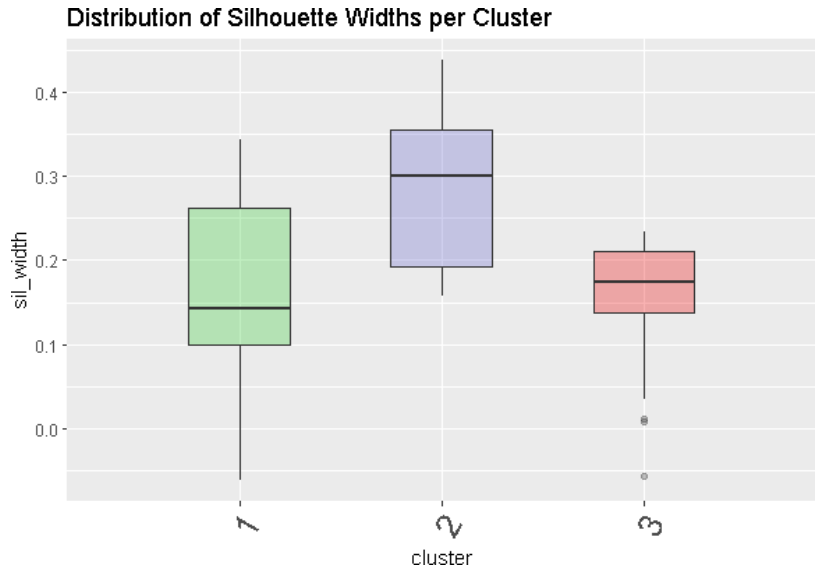

Figure 6.13: Distribution of Silhouette widths per cluster

Figure 6.14: Silhouette widths within cluster

From the above figure it can be noticed that Cluster 1 and 3 contain some outliers.

Final distribution of clusters of first 10 values are shown below in a table:

Cluster	Age_Numeric	Cloth	Gadget	Books	Plants
1	2	3	3	3	1
1	2	3	3	3	1
2	1	3	1	3	2
1	2	3	2	4	1
2	1	3	3	1	1
1	1	3	4	2	1
1	4	3	4	3	1
1	5	3	4	2	2
1	2	3	3	3	5
2	2	3	1	2	2
2	1	3	1	3	1
2	2	3	1	2	1
1	3	3	4	4	6
1	2	3	3	3	2
2	3	1	1	2	1
1	2	3	5	5	1
1	2	3	4	3	1
1	2	3	3	2	2

Table 6.1: Final distribution of clustering

6.1.2 Classification using SVM (Supervised Learning):

For this project, the Support Vector Machine (SVM) algorithm was utilized due to its robustness and relatively straightforward principles. This algorithm uses a hyperplane to distinguish between data points with the greatest margin possible, which is why it's also referred to as a discriminative classifier. By identifying an ideal hyperplane, SVM can effectively classify new data points.

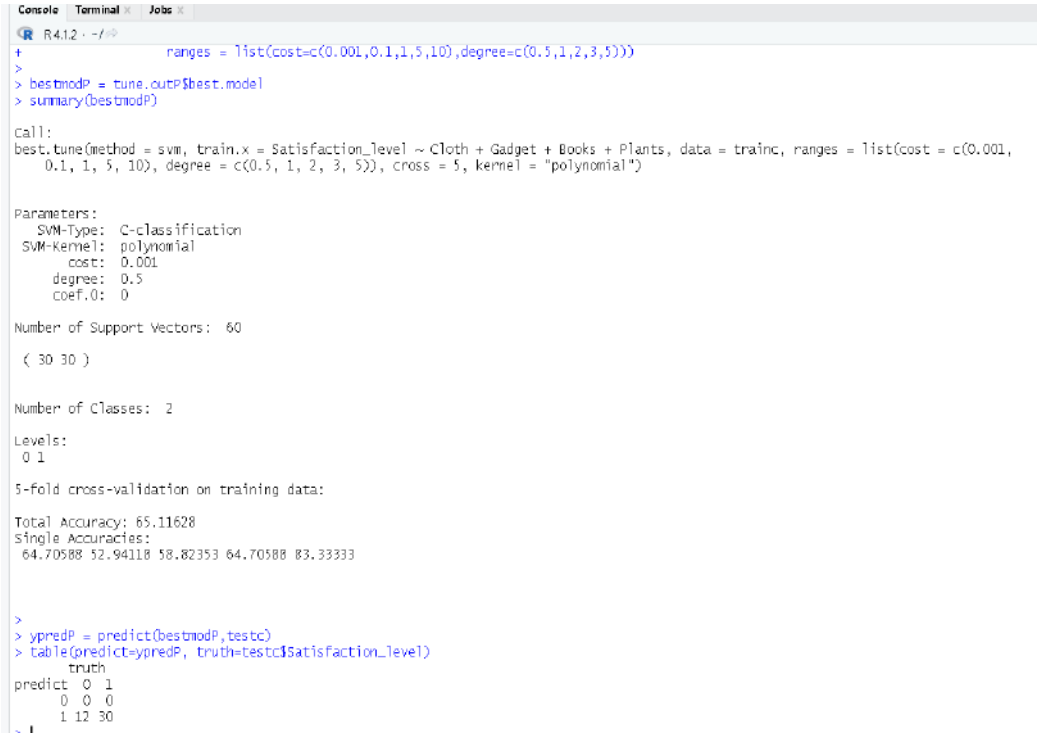

Figure 6.15: 5-fold cross validation output (Polynomial Kernel)

The polynomial kernel is a type of kernel function that is frequently utilized in machine learning with models like support vector machines (SVMs). Its purpose is to capture the similarity of vectors (training samples) in a feature space by raising the original variables to different powers, enabling the model to learn non-linear relationships between variables. In this work, by making the customer satisfaction level as dependent variable, we estimated it with the help of independent variables like cloths, gadgets, books, and plants. For the single level folds, the accuracy levels came out about 64.7%, 52.94%, 58.823%, 64.7%, and 83.33% consecutively. The number of classes is 2. In this work, we got the best result with radial

kernel of SVM. However, while using polynomial kernel, we used a degree of 0.5, cost value of 0.001, and the number of support vectors in this case was 60 for both levels (0 and 1).

```
>
> svmfit = svm (Satisfaction_level~Cloth+Gadget+Books+Plants , data=trainc,kernel="radial",cost=10,scale=TRUE)
> summary(svmfit)

Call:
svm(formula = Satisfaction_level ~ Cloth + Gadget + Books + Plants, data = trainc, kernel = "radial", cost = 10,
    scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  10

Number of Support Vectors:  73

 ( 30 43 )


Number of Classes:  2

Levels:
 0 1



>
>
>
> ypredict=  predict(svmfit,testc)
> table(predict=ypredict, truth= testc$Satisfaction_level)
       truth
predict  0  1
      0  6  5
      1  6 25
>
> svmfit$index
 [1]  1  3  5  7 12 16 19 25 26 29 32 34 39 41 52 58 59 60 61 65 66 68 69 73 76 79 81 83 85 86  2  4  6  8 11 13 15 18 20 21 22 24 27 31 33
[46] 35 36 37 38 40 42 45 46 47 48 49 50 51 53 54 55 57 62 63 64 70 71 74 75 78 80 82 84
>
```

Figure 6.16: Result using Radial Kernel Function

For this project purpose, radial and polynomial kernel functions have been used for data classification and prediction. 5-fold and 10-fold cross validation showed the best result and high prediction accuracy on the dataset. Polynomial kernel with 5 cross validations showed an accuracy level of 66%. Radial kernel function showed an accuracy level of 73% to perfectly predict whether a customer is highly satisfied or not by analysing that person's preference levels. Train data set size was two thirds of total data.

CHAPTER 7

CONCLUSION AND RECOMMENDATION

7.1 Conclusion:

The purchasing nature or behaviour of consumer would always be complex to predict. Because human minds act differently in different situation. However, dealing with possible accurate and bigger data can bring result that is close to reality.

The purposes of this project was to come up with such accurate result as much as possible. And we used survey method to have as much real data as possible. Later on different clustering algorithms were used to figure out the customers who show similar buying nature.

Both the unsupervised and supervised algorithms worked quite well in the real world survey dataset. Though, the unsupervised algorithm suggested three clusters to classify the total dataset, the supervised did a great job while identifying satisfied customers on the basis of buying pattern of products.

But even after choosing survey method, all the data were not clean or true. Some people didn't answer all the question of the survey. Like, in the survey, for question number 5 (What type of products do you generally look for/buy from Facebook based business page?) 72 people among the 139 participants claimed that they purchase that "accessories" from Facebook. But they didn't provide enough information with the questions of price range or their comfort of purchasing accessories from F-commerce site. From those 72 voters some later on also mention they do not purchase products from Facebook at all. Such duality in opinion may cause faulty result and hence despite of having approximately 52% votes that entire category was omitted later on.

A second segment was also designed in the survey for F-commerce business owners. Unfortunately a very few people claimed to be the owner. Hence that entire segment was a complete waste and not used.

Also, from the 139 participants only 128 were originally F-commerce users as well as customers. It is inevitable that, for most machine learning model the more the data input is, the better or accurate the result will be. But in this advanced world, when there remain tight competition in every business including F-commerce, one might want to know the customer

segmentation for his very own product and serve those niche market only. And for that surveying with a targeted group of customers, collecting true and clear data and then acknowledging the future customer beforehand and showing relevant advertisement to relevant consumers will be a smart move.

7.2 Recommendation:

Future work scope generates from the drawback of any present or existing model. And like any other forecast model, this project also has its own error.

Overcoming these limitations can be a scope to work further on this very model in the near future. One of the biggest constraints of this proposed system faced was that, due to the virtual survey, some people have provided biased information. Such untrue, vague information or nominal value leads to faulty numeric value. While clustering with such diversified value the data automatically overlaps more than it actually should; concluding faulty results and eventually becoming another limitation.

A good number of unfiltered or unbiased data and a field survey can make this model more accurate in order to predict F-commerce users buying behaviour beforehand. Moreover, collected data from F-commerce experts and personal business owner can deliberately provide this model with better accuracy in results.

Also, while completing this survey, some of the participants appeared as anonymous. Here the scope of this project also increases because, if it were possible to get the original Facebook ids or maybe email ids of all of those participants, this model would have only target those customers for digital market advertisement or campaign. And that would reduce the marketing campaign cost.

So, it is possible to work on this model further with a bigger but true data set and targeting the possible consumers more precisely by having their social media links (with their sole permission obviously) or email ids.

Sub-segmentation of product categories (book: genre, clothe: fabric etc.) can also be done in future for even more precise prediction.

REFERENCES

Chang, S., Zhang, Y., Zhang, Y., Tang, J., Yin, D., Chang, Y., . . . Huang, T. S. (2017). Streaming Recommender Systems. *26th International Conference on World Wide Web* (pp. 381–389). Perth, Australia: International World Wide Web Conferences Steering Committee. doi:10.1145/3038912.3052627

Géron, A. (2017). Support Vector Machine. In A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow* (p. 145). O'Reilly Media, Inc.

Géron, A. (2017). Principal Component Analysis. In A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow* (p. 211). O'Reilly Media, Inc.

Gupta, R., & Pathak, C. (2014). A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing. *Procedia Computer Science, 36,* 599-605. doi:10.1016/j.procs.2014.09.060

Hu, D. J., Hall, R., & Attenberg, J. (2014). Style in the long tail: discovering unique interests with latent variable models in large scale social E-commerce. *20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1640–1649). New York, New York, USA: Association for Computing Machinery. doi:10.1145/2623330.2623338

Huang, C., Wu, X., Zhang, X., Zhang, C., Zhao, J., Yin, D., & Chawla, N. V. (2019). Online Purchase Prediction via Multi-Scale Modeling of Behavior Dynamics. *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2613–2622). AK, Anchorage, USA: Association for Computing Machinery. doi:10.1145/3292500.3330790

Jagiela, A. R., Ven, V. v., Knöchel, V. O., Uhlhaas, P. J., Vogeley, K., & Linden, D. E. (2010). Resting-state functional network correlates of psychotic symptoms in schizophrenia. *Schizophrenia Research, 117*(1), 21-30. doi:10.1016/j.schres.2010.01.001

Jia, R., Li, R., Yu, M., & Wang, S. (2017). E-commerce purchase prediction approach by user behavior data. *International Conference on Computer, Information and Telecommunication Systems* (pp. 1-5). Dalian, China: IEEE. doi:10.1109/CITS.2017.8035294

Liu, H., & Cocea, M. (2018). Fuzzy rule based systems for gender classification from blog data. *Tenth International Conference on Advanced Computational Intelligence* (pp. 79-84). Xiamen, China: IEEE. doi:10.1109/ICACI.2018.8377585.

Martino, F. D., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage, 43*(1), 44-58. doi:10.1016/j.neuroimage.2008.06.037

Mikolov, T. K. (2010). Recurrent neural network based language model. *Interspeech*, 1045-1048.

Mondal, B., & Choudhury, J. P. (2013). A comparative study on k means and pam algorithm using physical characters of different varieties of mango in india. *International Journal of Computer Applications*, 21-24. doi:10.5120/13485-1189

Niklas, R., Outi, S., & Mikko, S. (2013). Predicting purchase decision: The role of hemispheric asymmetry over the frontal cortex. *Journal of Neuroscience, Psychology, and Economics,, 6*(1), 1–13. doi:10.1037/a0029949

Pan, B., Demiryurek, U., & Shahabi, C. (2012). Utilizing Real-World Transportation Data for Accurate Traffic Prediction. *IEEE 12th International Conference on Data Mining* (pp. 595-604). IEEE. doi:10.1109/ICDM.2012.52

S.Dixon. (2022). *Leading countries based on Facebook audience size as of January 2022.* Statista. Retrieved from https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/

Sarwar, D., & Deb, P. (2022). Multi-Criteria Stochastic Inventory Classification, Optimization & Simulation of an E-commerce Site by Integrating Support Vector Machine & Genetic Algorithm. *7th North American Conference on Industrial Engineering and Operations Management.* Orlando, Florida, USA.

Shahriar, R. (2021). *F-commerce in Bangladesh: Problems and Prospects.* Bangladesh: United News of Bangladesh. Retrieved from United : https://unb.com.bd/category/business/f-commerce-in-bangladesh-problems-and-prospects/77360

Sharma, R. (2022, July 27). *upGrad.* Retrieved from www.upgrad.com: https://www.upgrad.com/blog/clustering-and-types-of-clustering-methods/

Tyagi, N. (2021). *5 Clustering Methods and Applications.* Analytics Step.

Wang, Y., Chattaraman, V., Kim, H., & Deshpande, G. (2015). Predicting Purchase Decisions Based on Spatio-Temporal Functional MRI Features Using Machine Learning. *Transactions on Autonomous Mental Development, 7*(3), 248-255. doi:10.1109/TAMD.2015.2434733.

Wu, S., Ren, W., Yu, C., Chen, G., Zhang, D., & Zhu, J. (2016). Personal recommendation using deep recurrent neural networks in NetEase. *32nd International Conference on Data Engineering (ICDE)* (pp. 1218-1229). Helsinki, Finland: IEEE. doi:10.1109/ICDE.2016.7498326

Yi, Z., Wang, D., Hu, K., & Li, Q. (2015). Purchase Behavior Prediction in M-Commerce with an Optimized Sampling Methods. *IEEE International Conference on Data Mining Workshop* (pp. 1085-1092). Atlantic City, NJ, USA: IEEE. doi:10.1109/ICDMW.2015.69

Yousef, M., Jung, S., Showe, L. C., & Showe, M. K. (2007). Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data. *BMC Bioinformatics, 8*(1), 1-12. doi:10.1186/1471-2105-8-144

Zhang, Y., & Pennacchiotti, M. (2013). Predicting purchase behaviors from social media. *Proceedings of the 22nd international conference on World Wide Web* (pp. 1521–1532). Rio de Janeiro, Brazil: Association for Computing Machinery. doi:10.1145/2488388

Zhao, Q., Zhang, Y., Zhang, Y., & Friedman, D. (2017). Multi-Product Utility Maximization for Economic Recommendation. *Tenth ACM International Conference on Web Search and Data Mining* (pp. 435–443). Cambridge, United Kingdom: Association for Computing Machinery. doi:10.1145/3018661.3018674

Zhao, Y., Yao, L., & Zhang, Y. (2016). Purchase prediction using Tmall-specific features. *Concurrency and computation: practice and experience*, 3879-3894. doi:10.1002/cpe.3720

Zhou, M., Ding, Z., Tang, J., & Yin, D. (2018). Micro Behaviors: A New Perspective in E-commerce Recommender Systems. *The Eleventh ACM International Conference on Web Search and Data Mining* (pp. 727–735). CA, Marina Del Rey, USA: Association for Computing Machinery. doi:10.1145/3159652.3159671