Market Sales Prospecting By Analyzing Customer Buying Pattern Using Machine Learning

By

Tanjil Ahmed
14101053
Salman Rahman
15301021
Niloy Routh
15301015
Eftakhar Alam Nirob
15201036

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
August 2019

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Salman Rahman 15301021	**Tanjil Ahmed** 14101053
Niloy Routh 15301015	**Eftakhar Alam Nirob** 15201036

Approval

The thesis titled "Market sales prospecting by analyzing customer buying pattern using machine learning" submitted by

1. Tanjil Ahmed (14101053)
2. Salman Rahman (15301021)
3. Niloy Routh (15301015)
4. Eftakhar Alam Nirob (15201036)

of Summer, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on August 28, 2019.

Examining Committee:

Supervisor:
(Member)

Dr. Muhammad Iqbal Hossain
Assistant Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
BRAC University

Departmental Head:
(Chair)

Dr. Mahbubul Alam Majumdar
Professor
Department of Computer Science and Engineering
BRAC University

Acknowledgement

First of all, we would like to praise the Almighty for whom our thesis has been completed without any major hurdle.
Secondly, we would like to thanks our advisor Dr. Muhammad Iqbal Hossain for his kind support and guidance throughout the journey. He helped us with whatever we asked of him.
And finally, to our parents without their thorough support, it may not be possible. With their kind support and prayer, we are now on the verge of our graduation.

Abstract

For improving business, the organization needs to analyze the buying pattern of their customer to keep track of all products that are being sold the most so that they could keep the stocks of those products and remove those which sells the least. Analyzing cross-selling, up-selling of products is one of the major issues for identifying the buying frequency pattern. We have proposed a machine learning-based model to recommend and predict through K-Nearest Neighbor (KNN), fuzzy KNN (fKNN) Single Value Decomposition (SVD) algorithm to compare the best outcome. Our proposed recommendation model can be used to get an idea of which products need to keep more on their shelves and which products not for the convenience of the customers. Moreover, our proposed predictive model will help to forecast future profitability with which business developer will be better equipped to see the bigger picture. So, we are hopeful our model will help to serve the interest of all the stakeholders.

Keywords: Cross-selling, Up-selling, KNN, fKNN, SVD, Recommendation, Forecasting.

Contents

Declaration ii

Approval iii

Acknowledgement iv

Abstract v

List of Figures viii

List of Tables ix

Nomenclature x

1 Introduction 1
- 1.1 Motivation . . . 2
- 1.2 Problem Statement . . . 2
- 1.3 Objective and Contributions . . . 3
- 1.4 Thesis Structure . . . 3

2 Background 4
- 2.1 Product Recommendation . . . 4
- 2.2 Product Prediction . . . 5
- 2.3 Cross and up-selling . . . 6

3 Related Works 7
- 3.1 Literature Review . . . 7
- 3.2 Predictive model . . . 9
- 3.3 Recommendation Systems . . . 11
  - 3.3.1 KNN . . . 11
  - 3.3.2 KNN fuzzy . . . 12
  - 3.3.3 Singular Value Decomposition (SVD) . . . 14

4 Dataset Description 16

5 Experimental Results 23
- 5.1 Result for Recommendation . . . 23
- 5.2 Result for Prediction . . . 27

6 Result Analysis 31

7 Conclusion 33

References **34**

List of Figures

Figure 1 Cross and Up-selling. 6
Figure 2 A Carbon Dioxide against Year by year prediction. 10
Figure 3 Scattered Representation of Time Series. 10
Figure 4 Visual Representation of KNN. 11
Figure 5 Predicted vs. Observed Wavelength. 14
Figure 6 Comparison between 2 SVDs. 15
Figure 7 Heatmap for first 20 products. 22
Figure 8 Future Profits for Phones (X=Years, Y=Profits in USD). . . . 28
Figure 9 Future Profits for Phones and Papers (X=Years, Y=Profits in USD). (b) Future Profits for Phones and Accessories (X=Years, Y=Profits in USD), (c) Future Profits for Phones and Chairs (X=Years, Y=Profits in USD) (d) Future Profits for Phones and Fasteners. . . . 30

List of Tables

Table 1 Pre-processed Dataset 17
Table 2 Dataset of Global superstore 18
Table 3 Dataset of Global superstore 19
Table 4 Dataset of Global superstore 20
Table 5 Recommended Products for KNN & FKNN 24
Table 6 Recommended Products for SVD 25
Table 7 Sample Cross-sell output for KNN, FKNN and SVD 26
Table 8 Sample Up-sell outputs for KNN, FKNN and SVD. 27
Table 9 Recommended product list. 28
Table 10 Precision, Recall, MAE and RMSE of KNN, FKNN, SVD. . . 32

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document.

ARIMA	Autoregressive integrated moving average
KNN	K-Nearest Neighbor
fKNN	Fuzzy K-Nearest Neighbor
SVD	Singular Value Decomposition

Chapter 1

Introduction

Business plays a big role in a country's economy and many developments are being made to make businesses more profitable and efficient as much as possible. Scholars, researchers, analyst have come up with many ways in which one can improve their business. Customers are the lifelines of any business, meeting their desired requirements are the key to a successful business. Relationship marketing and customer-centric selling strategies have become Ubiquitous as sales organizations have realized the benefits of transitioning their strategies towards relationship development and customer solutions. The better an organization understands the relationship between them and customers, the better their sales will be.

One way of finding out what customers really want is to analyze a customer's buying pattern, it implies that what a customer but for a certain period of time. Basically, it is evaluating a buyer's past product purchases to predict their future buying pattern, this way they can evaluate the common buying patterns among the customers. It will help a business organization keep track of what all products that are being sold the most so that they could keep the stocks of those products and remove those which sells the least. In recent years, the debate on cross-selling up-selling has emerged as a new concept for increasing sales effectiveness. Cross-selling means to sell a different product or service to a customer. Up-selling means to persuade a customer to buy something additional or more expensive. Most of the customers do not buy only one product, they buy two or more products at once. So analyzing them will give us an idea which products do a customer usually buy together from there we can draw a conclusion that what products should go together in a cross-sell.

When a customer buys a product, we can suggest him a better version of the product which might be a little expensive than the one he is currently looking at. A customer might not know that with just getting a little cash he might get a better version of the product he currently willing to buy. Product recommendation is an analytical process of finding out which product sells were the most depending on the season. Cross-selling can boost your overall sales if it is implemented in the right way, it may bring bizarre items together for example Baby diapers beer, etc. Upselling has also its share of advantage, it may recommend product variant that many did not know even existed.

Businesses do not become successful overnight, they need a lot of analyzing, adaptability if they want to stay in the market and compete because it is a fast-changing world we live in. Every minute, every second a consumer's needs desires are changing, today he might want something the very next day he might want totally something different, something totally out of the blue. So, a business should keep up with the fast-paced world and the ever-growing competitors. If a customer does not get what

he wants from you, he will move on to the next seller because everyday competitions are entering the market if you want to stay ahead of them then you have to know what your customers want and when do they want it if you do not want to lose them to your competitors.

1.1 Motivation

While starting research on up cross-selling, we noticed that little amount of work has been done in the past on these topics, very few comprehensive works were published on these topics. We thought that as less research had been done on these topics, our field of research in it could be broad, bring in a new perspective in it, see where it takes us must manipulating various parameters. In this paper, we lay a foundation for future works on these topics by doing comprehensive research given that very few papers had been published on this topic. Bound by limitations, we might not be able to develop it as much as we want but we might take it to a point where in the future, realizing the potential cross-selling up-selling has on sales, the researchers will take an interest in our work and will pick up right from we left it off. In this digital world very few fields are still left untouched especially in the business fields no one even cares. We wanted to research on a field where very few works had been done in the past.

1.2 Problem Statement

We know that problem recognition is the first step towards solving a problem. In market economies, identifying customers' demand plays a pivotal role to produce or generate the desired service and subsequent market penetration. Cross-selling and up selling technique have been used for long to attract customers on the bundle offers of their products. But a huge chunk of customers could not find their preferred bundle from hundreds of products due to some limitation of précised calculation. The sellers also find it harder to maintain the inventory due to lack of proper understanding of demand on this particular sphere. Hence, we applied some algorithm in calculating which of the products best suited for cross bundling. Thus, it should have helped customer to go for their favorite bundle amongst myriad of items. Additionally, it would also make the seller happy since it would give them a clear picture of which items should be picked with a view to having more profits and revenue.

1.3 Objective and Contributions

.

- Applied KNN, fKNN and SVD for product recommendation
- Applied KNN, fKNN and SVD for product recommendation
- Comparative analysis.
- Adding new dimension for further exploration.
- Change of output by varying the parameters.
- Pleasing both buyers and sellers.

1.4 Thesis Structure

In chapter 2 we described what is meant by product prediction, product recommendation cross up-sell. After that chapter 3 covers the information about the previous works done related to our research the algorithms we used in our research. Then in chapter 4 we discuss the dataset and various aspect of dataset attributes. Chapter 5 shows the separate research results for recommendation prediction. In chapter 6 we analysis of our Results and finally in the chapter 7 we conclude our paper and discuss future direction.

Chapter 2

Background

For clear understanding, we divided background analysis in product recommendation, product prediction and up-sell/cross-sell.

2.1 Product Recommendation

At present, the customer engaged in the multidimensional role in the decision-making process over a company or organization's strategy. Even the retailers want to know the deep understanding of the shopping basket a customer carrying. Since it will help them to formulate the strategy of selling. Recommender systems are one of the prime technological revolutions that give the seller detailed insights about the customers' preference.Recommender systems solve this problem by searching through large volume of dynamically generated information to pro-vide users with personalized content and services [23]. Targeted cross-selling procedures can be boosted through the efficient procedure of this recommender system which makes establishing customer's loyalty and fulfilling their needs. Usually, this recommender system can be classified by two methods.
First one the content-based recommendation deals with similar products to those the customer has purchased in the past. Another one the collaborative filtering makes a recommendation based on the products owned by a user whose preference is quite similar to those given user. The hybrid approaches of both methods are generally used to complement one another off late. Generally, two sources of information are used in recommender system. Using the clustering to identify similar groups of the customer based on their prior purchasing pattern. Another one works with the concept of performing rule association mining on the same basket of the cluster to extract the relationship amongst the products.
As this relationship is derived from the purchases of similar group of customers that have purchased at least one product in the same time period. So, it is expected that additional relationship of the products must be formed that are not derived from only the customer's past baskets. In short, the recommender system formulates the collaborative filtering keeping the idea of content-based recommendation. The target of the recommender system is to provide periodically relevant and personalized items to loyal customers to raise the attachment of the customers. On the contrary, the recommendation of no interest to the consumer or excessive recommendation could have adverse effect Poor recommendation can trigger the characteristic errors like false negative which means that a product liked by the customer but not recommended and false-positive which are products that are recommended in the first place but not liked by the customer. The false-positive certainly lead to incurring more losses by losing the customer. To avoid such scenario the carefully planned recommender

system for the target customer who is going to buy that recommended products is vital. So recommendation should be specific to each of the clusters and having large numbers of customers inside the cluster can be helpful to select the target market.

2.2 Product Prediction

In this highly competitive world, people would do just about anything to stay ahead of others. Competition is present in every aspect of life from being born to doing business. Doing business in this era of advanced technology is really tough; you not only have to keep up with the latest technology that comes out literally every day but also with the ever-lasting changing demands of the customers. With technology their demands are also changing every day. To stay in business and also ahead of other businesses and on top of customers' demands, is to predict their demand for the long term. It will not be possible to predict their short-term demand but predicting their long-term ones could give them an invaluable edge over other businesses. Predicting might save businesses from incurring losses and going out of market. Many algorithms are available for doing such predictions.
Predictions are not done only on customers, demands but also on product sales like which products are likely to sell the most in the future and how much profit shall a business get from that sale. It can be done based on many criteria like age, region, and season etc. Not only does prediction helps a business stay afloat but also it helps to meet customers' desires. Producers will only produce the products which will be in demand or might be in the near future. The forecasting methods mainly look for a trend or (seasonal) pattern in the historical sales data and sometimes relate this to events of other source [24]. Our paper tries to compare algorithms to see which one gives the best accuracy. Although predictions might not also be correct at least it can do just enough to keep the businesses running. Predictions can be tough if you try to cover all the sectors at once. To get the best out of it, it's best to do it base on the location, type of your business and the type of customers that usually buys from you. We tried to cover in our paper predictions based on region and season. You might also want to take into consideration the season as it also might have an effect on your sales. In this fast-moving world, the businesses which can adapt quickly to customers' needs will only survive and the rest will perish with time. Although the findings of our research would not be as extensive or accurate as we would like to be because of our limitation of knowledge and technology. Here, we used the technic to predict how the to cross-sell can change over a course of time. It will depend on region and season. Cross-sell might be a huge benefactor from this as producers will know by seeing a pattern of purchase what products customers usually buy together. Using that data businesses can put that product together for a lower price to attract more customers. In this way, both the customers and the sellers will be happy as businesses are earning more money and consumers are getting their desired products together for a cheaper price. Product prediction has a huge potential in this generation of technology.

2.3 Cross and up-selling

Cross-selling is a strategy to encourage the customer to buy complementary products with the existing item where the complementary products are from different category. For instance in figure 1, if a customer who is keen to purchase a burger is offered French fries then it is considered a cross-selling strategy. On the other hand, upselling indicates encouraging customers to buy upgraded or expensive version of the products that are initially chosen [8]. For example, a customer is interested to buy a Smartphone preferring the configuration of 4 GB Ram and 64 GB Rom then if the seller convinces him to purchase the Smartphone's more expensive and better version having 6 GB Ram and 128 GB Rom then this will be termed upselling. Both cross-selling and upselling are closely related since both of the strategies emphasizes exploring the additional value to customers. So, the proper understanding of the relation between what the customer prioritizes most and in respond to what the offers the sellers put before them is crucial in order to sustain the customer in the long run [25]. It is quite understandable that targeting the existing customer for upselling and cross-selling is lot easier than to a new customer base. Moreover, it helps to build a more lasting relationship with the customer due to the enhanced value derivation out of their purchase, which makes their job easier. Apparently, the increasing Customer Lifetime Value (CLV) is another significant factor in this context. Exploring the market basket analysis and customer buying behavior is the key to accurately predict the items that customer likely to purchase.

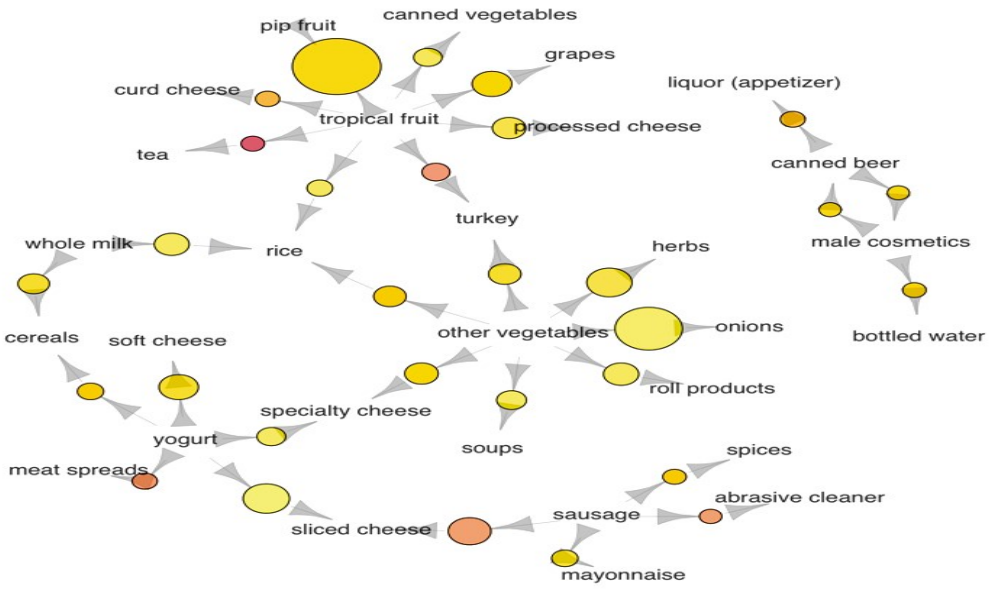

Figure 1: Cross and Up-selling.

Chapter 3

Related Works

3.1 Literature Review

Many researches have been done in this field. In [1] authors used cross-selling to establish a relationship between a customer and an organization to recognize the importance of a loyal customer to an organization. They used up-selling to upgrade or improve the conditions of a product already purchased. Furthermore, [2] Suggests that cross-selling is higher when made face-to-face interactions rather than buying online not only that Cross-selling involves the selling of peripheral services as well as 'upselling', that is, the selling of more expensive services to the customers which also increases because of face-to-face interactions. Another research paper[3] suggests that observation of cross-selling and up-selling traditionally rely on transactional databases which do not assess the salesperson's orientations and attitudes To overcome this limitation, the authors capture the behavioral tendencies towards cross-selling and up-selling by salespeople and embed them within a motivation-opportunity-ability (MOA) theoretical framework. An MOA framework is utilized to theoretically select and hypothesize variables which contingently influence the effects of salesperson cross-selling and up-selling on their performance and job satisfaction. This model can also be applied to understand a customer's psychology and what he wants.[4] states that cross-selling has evolved into a strategy for customer relationship management. The author divides the cross-selling into 2 groups: *Acquisition Pattern Analysis* and *Collaborative Filtering*, both the methods are important what the customer wants next. These analytical tools are important to make cross-selling work.
Moreover, [5] implies that the environment in which a selling team works plays a vital role in making the cross/up-selling successful. It further states that working in a team rather than alone deals better with technological complexity in products and services, supplier rationalization, coordinated buying, and greater customer expectations, it also improves internal co-ordination. Both the sell techniques it says fail most of the time because firms overlook the importance of a selling team. [6] argues that cross up-selling methods can also be used in banking sectors but the author warns that this could be particularly difficult because companies have different products operate under a complex set of business constraints. Data mining could be used to help with some of the complexities as we can collect data of customers that contains an expected profit of each product for each customer. The ideal approach is yet another way to make things easier in the banking sector. If the above methods fail then the company could try the practical approach so they can experience first-hand what was holding them back in the first place. In [7] the authors find a unique way of predicting cross-sell using radio frequency technology. They developed 2 systems by using radio frequency that is: Smart dressing system intelligent product cross-selling system. The Smart dressing system is designed based on the current frequency allocation

for RFID assigned by The Office of Telecommunication Authority (OFTA) of Hong Kong, and the frequency band is 920–925 MHz The back-end subsystem, i.e. Intelligent Product Cross-selling System (IPCS), is designed to assist fashion designers or stylists to streamline the process in making the mix-and-match pairs.
After the successful test run of these systems, it was implemented in some major market chains of Hong Kong and it really proved to be successful and it saw an increase of 20 % in sales. Cross-selling is a strategy to encourage the customer to buy complementary products with the existing item where the complementary products are from different category. For instance, if a customer who is keen to purchase burger is offered French fries then it is considered a cross-selling strategy. On the other hand, up selling indicates encouraging customers to buy upgraded or expensive version of the products that are initially chosen. For example, a customer is interested to buy a Smartphone preferring the configuration of 4 GB Ram and 64 GB Rom then if the seller convinces him to purchase the Smartphone's more expensive and better version having 6 GB Ram and 128 GB Rom then this will be termed up selling. Both cross-selling and up-selling are related since both strategies emphasizes on the exploring the additional value to customers. Cross-selling and up-selling are implemented with a vision of retaining the existing customer. In order to extract more revenue from the existing customer pool cross and up-selling has been a proven game-changing tools over the years across the globe.[8] It has been found in researches that acquisition of a new customer is 5-25 times more expensive than retaining a new customer. Consequently, a retained customer will definitely spend more and purchase more frequently [9]. Chances are 60-70 % of selling a product to an existing customer compared to 5-20 % to a new customer. So it is quite understandable that any company which is not cross and up-selling they are losing the leverage by leaving money on the table. The cornerstone of a successful packaging like cross and up-sell is to know the appropriate offer to be placed. Since the outcome of a successful cross and up-selling will ensure the higher conversion rates ,boost the number of daily orders and expose various categories from the inventory (best seller, items with better reviews, outsiders) the resources need to be optimized for coming out of a tempting offer from a customer's perspective.
Aligning the up-sell model according to the customers wants is crucial [10]. For example a customer who has got interested in a sedan car at a price of $20k more likely not to be inclined towards a $50k sports car. Rather he might be scale up his range for a $23k sedan car [11]. Amazon attributes 35% of its total revenue to cross-selling by implying some basic strategy like promoting sections like "Customer who bought this item also bought ", "Frequently bought item " inside the windows viewers currently looking for similar products. People get influenced with such tagline frequently and thereby ordering the recommended items for them. So, to sum up all these, it can be said that a lot of research has been done in our field with coming up with new technologies to predict them but none of came as close as our research most of the research work had only been focused on cross-selling as it is the tougher of the two. Our research has focused equally on both cross up-selling to accurately measure

which algorithms goes better with the 2. If implemented properly our research could do businesses a great help.

3.2 Predictive model

A period arrangement is where measurement is recorded over normal time interims. Contingent upon the recurrence, a period arrangement can be of yearly, quarterly, month to month, week after week, day by day, hourly, minutes and even seconds shrewd as shown in figure 2. Time arrangement estimating predicts future information focuses dependent on watched information over a period known as the lead-time in figure 3. The motivation behind estimating information focuses is to give a premise to financial arranging, generation arranging, creation control and improving modern procedures. The real target is to acquire the best estimate work, i.e., to guarantee that the mean square of the deviation between the real and the determined qualities is as little as feasible for each lead-time. Much exertion has been committed in the course of recent decades to the advancement and improvement of time arrangement estimating models. Conventional models for time arrangement determining, for example, the Box-Jenkins or autoregressive coordinated moving normal (ARIMA) model, accept that the considered time arrangement is created from straight procedures [19]. Time arrangement anticipating is a multidisciplinary logical instrument used to take care of expectation issues. Its execution is simple and adaptable on the grounds that it just requires authentic perceptions of the vital factors. ARIMA was first exhibited by Box and Jenkins in 1976. The general condition of progressive contrasts at the dth distinction of Xt is as per the following [20]:

$$\triangle^d X_t = (1-B)^d X_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (i)$$

Where d is the difference order and is usually 1 or 2, and B is the backshift operator. The successive difference at one-time lag equals to,

$$\triangle^1 X_t = (1-B)X_t = X_t - X_{t-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (ii)$$

An ARIMA model is characterized by 3 terms: p, d, and q where,
p is the order of the AR term
q is the order of the MA term
d is the number of differences required to make the time series stationary
In this work, the general ARIMA (p, d, and q) is briefly expressed as follows:

$$\Phi_p(B) W_t = \theta_q(B) e_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (iii)$$

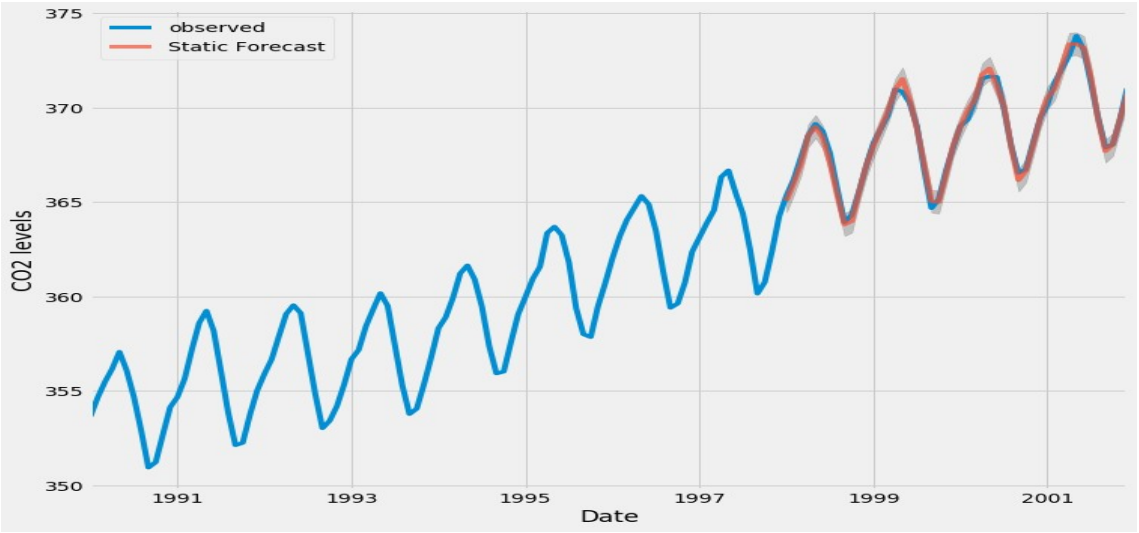

Figure 2: A Carbon Dioxide against Year by year prediction.

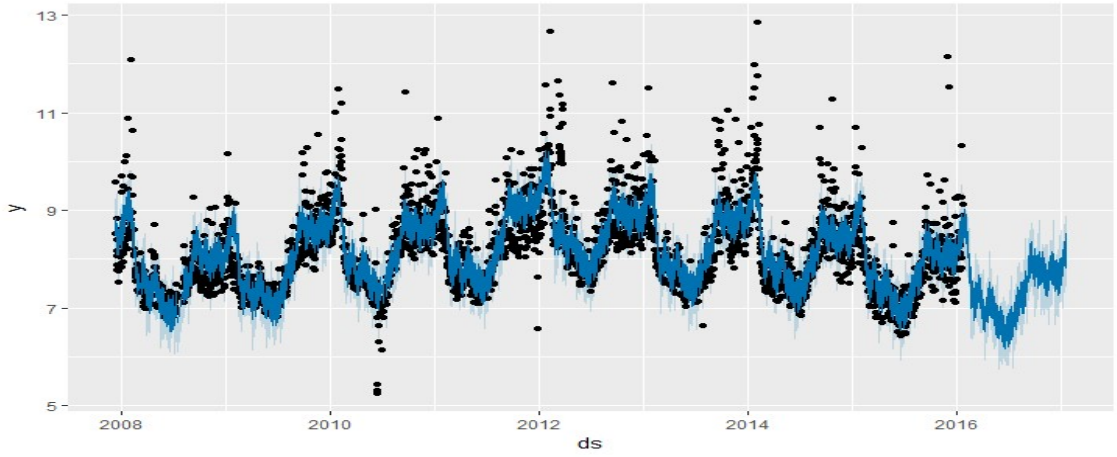

Figure 3: Scattered Representation of Time Series.

3.3 Recommendation Systems

3.3.1 KNN

K- Nearest neighbor is a supervised machine learning algorithm to implement both classification and regression related problem in an easier way. In machine learning a supervised data indicates labeled input data to make learn a function and based on the learning further classification of an unlabeled set of data can be derived[26]. For example, let's assume computer is a child and we need to make it learn how a cow looks like. So in this case, we have to give lots of images featuring cow with various angle and therefore termed them" A cow". We also provide some other animals images which are grossly termed "Not a cow" Then after the training session if we would show multiple pictures of cow and other animals and ask to term accordingly.If the child could term the images correctly then the whole process of learning can be compared with supervised data learning method. In K Nearest neighbor a test sample is provided as the class of the majority of its nearest neighbor. In plain words, if a random person similar to his neighbor he/she is one of the neighbours. In other words, if a lion is similar to tiger, cat, and leopard than an apple, orange and jackfruit then most likely lion is an animal.K-nearest neighbor algorithm defines the things according to the proximity, works out on the principle of sorting similar things which are placed closely in figure 4[12],

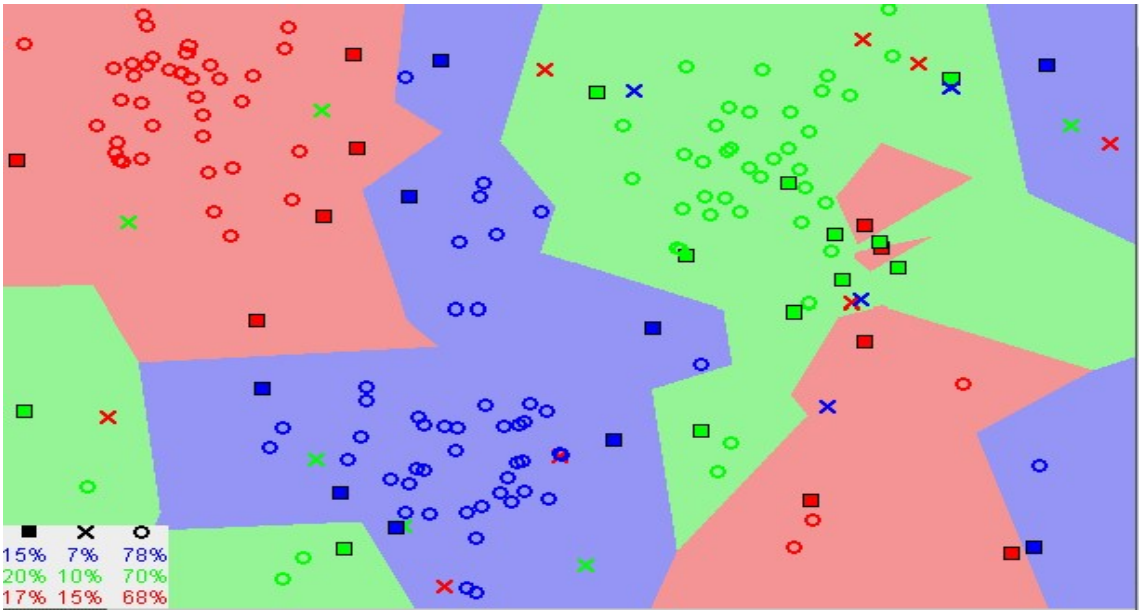

Figure 4: Visual Representation of KNN.

We can easily notice that similar data points are placed with each other by calculating the distance between the two adjacent points. There are some distance matrix exist to calculate amongst the Euclidean distance is quite familiar. The procedures lead to the k-nearest neighbor algorithm involves :
1) Loading the data
2) Initialize k to the chosen neighbor numbers
3) Calculate the distance between query and current examples in the data
4) Add distance (Euclidean) as well as index of the examples in an order(ascending or descending)
5) Pick first k entries from order
6) If regression needs to get extracted then the mean of k labels must be returned
7) If classification needs to get extracted then the mood of k labels must be returned
With the increasing value of k we can get smoother and more defined boundary while extracting classification. [13] Similarly, the more decreased value of K will result in a less stable prediction model. The overall accuracy of the k nearest lies on the increasing number of data points in the training set.

3.3.2 KNN fuzzy

The k-Nearest Neighbors (kNN) classifier is one of the best strategies in managed learning issues. It characterizes concealed cases contrasting their closeness and the preparation information, shown in figure 5. By the by, it provides for each named test a similar significance to characterize. There are a few ways to deal with upgrade its exactness, with the Fuzzy k-Nearest Neighbors (Fuzzy-kNN) classifier being among the best ones. Fluffy kNN processes a fluffy level of enrollment of each occurrence to the classes of the issue. Subsequently, it produces smoother outskirts between classes. Aside from the current kNN way to deal with handle huge datasets, there is anything but a fluffy variation to deal with that volume of data [14]. Grouping of items is a significant territory of research and application in an assortment of fields. Within the sight of full information of the basic probabilities, Bayes choice hypothesis gives ideal blunder rates. In those situations where this data is absent, numerous calculations utilize separation or closeness among tests as a method for arrangement.
The K-closest neighbor choice standard has regularly been utilized in these example acknowledgment issues. One of the challenges that emerge when using this method is that every one of the named tests is given equivalent significance in choosing the class participations of the example to be characterized, paying little mind to their 'commonality'. The hypothesis of fluffy sets is brought into the K-closest neighbor strategy to build up a fluffy rendition of the calculation. Three techniques for allotting fluffy enrollments to the named tests are proposed, and trial results and correlations with the fresh form are presented [15]. Some calculations dependent on fluffy set theory (FST) such as fluffy induction system (FIS) and versatile system based fluffy surmis-

ing system (ANFIS) have been achievement completely connected to noteworthy wave stature (SWH) forecast. Arrangement calculations are chiefly utilized for estimating the comparability of a lot of articles dependent on certain proportions of separation. The K-closest neighbor (KNN) calculation is one the most seasoned example classifier strategies with no preprocessing necessity. The choice principle of normal characterization calculations, for example, M5P, SVR and BN expect equivalent loads for article enrollment utilities, ignoring various examples of closeness. Exploiting fluffy set hypothesis, the FKNN has been appeared to not just have a lower blunder in ordering the articles yet additionally it has a more noteworthy certainty proportion of the grouping. The FKNN gives a progressively sensible vector of enrollment for the items and it additionally represents the level of article participation to the classes of objects [16].

As a straightforward, powerful and nonparametric order strategy, kNN calculation is broadly utilized in content arrangement. Be that as it may, there is an undeniable issue: when the thickness of preparing information is uneven it might diminish the exactness of order on the off chance that we just consider the arrangement of first k closest neighbors yet don't think about the distinctions of separations. To take care of this issue, we receive the hypothesis of fluffy sets, building another participation capacity dependent on archive likenesses. An examination between the proposed strategy and other existing kNN strategies is made by trials. The trial results demonstrate that the calculation dependent on the hypothesis of fluffy sets (fkNN) can advance the accuracy and review of content arrangement to a certain degree [17]. Fluffy arrangement is a generally investigated research arrangement of items in information sciences and building. With the range of time, it got new statures with huge upgrades as indicated by the necessities. Still there are a few issues to be examined and explained in a fluffy way; fluffy grouping of imbalanced information is one of them. Subsequently, the significance of fluffy closest neighbor came into the situation and conveyed in numerous applications. Different improved fresh closest neighbor methodologies are performing admirably on imbalanced informational collections, yet very little work has done on the fluffy closest neighbor for imbalanced information [18].

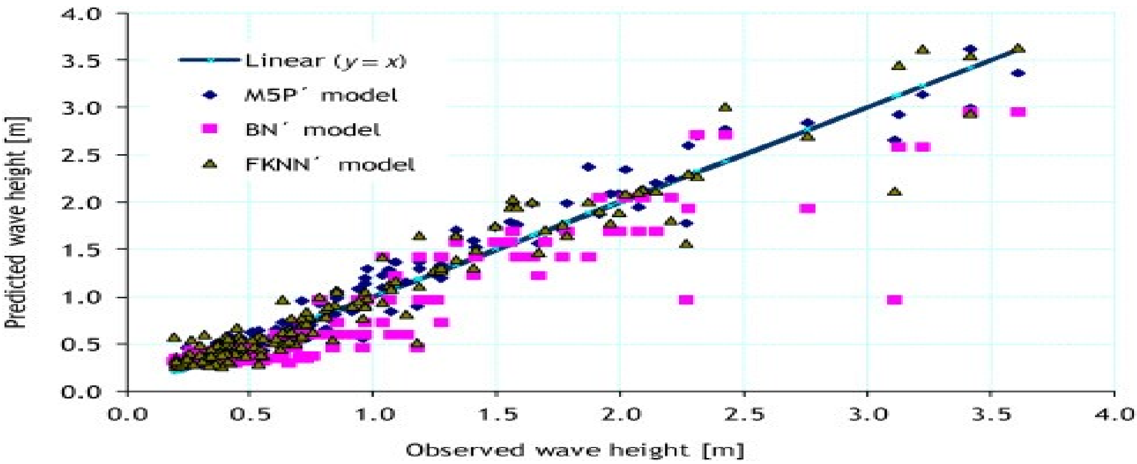

Figure 5: Predicted vs. Observed Wavelength.

3.3.3 Singular Value Decomposition (SVD)

The particular worth deterioration of a grid A_n is the factorization of A_n into the result of three networks $A = UDV^T$ where the sections of U and V are orthonormal and the lattice D is corner to corner with positive genuine passages[27]. In numerous applications, the information network is near a framework of low position and it is valuable to locate a low position lattice which is a decent guess to the information grid. We will demonstrate that from the solitary worth disintegration of A, we can get the lattice B of rank k which best approximates An visualized in figure 6 ; in actuality we can do this for each k. Additionally, solitary worth deterioration is characterized for all frameworks (rectangular or square) not at all like the more ordinarily utilized otherworldly disintegration in Linear Algebra [21]. On the off chance that grid A has a framework of eigenvectors P that isn't invertible; at that point A does not have Eigen disintegration. Be that as it may, in the event that an is a m x n genuine lattice with m > n, at that point A can be composed utilizing a purported particular worth deterioration of the structure

$$\text{A} = \text{UDV}^T \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (iv)$$

And

$$\text{U}^T U = I \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (v)$$

And

$$\text{V}^T V = I \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (vi)$$

(Where the two identity matrices may have different dimensions), and D has entries only along the diagonal.
For a complex matrix A, the singular value decomposition is decomposition into the

form

$$\text{A} = \text{U D V}^{H} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(vii)$$

Where U and V are unitary matrices, V^H is the conjugate transpose of V, and D is a diagonal matrix whose elements are the singular values of the original matrix. If A is a complex matrix, then there always exists such decomposition with positive singular values [22].

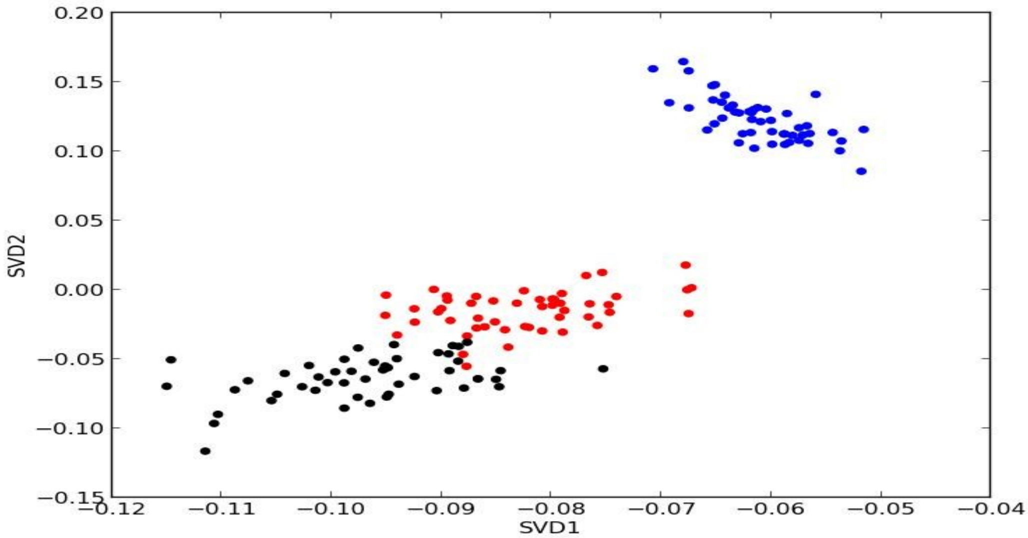

Figure 6: Comparison between 2 SVDs.

Chapter 4

Dataset Description

The Dataset was taken from a forum called Tableau community forums for global superstore 2016 containing more than 50000 purchase data.

In our research we had used global super shop dataset which contained various data about super shop parches of many geographical area of this earth. The dataset contained 50000 parches history. Here, we got 17000 unique customers who were parching 3500 unique products. The products were from 3 categories and 40 sub-categories. We had pre-processed the data (shown in Table 1) for our convenience of implementation. Firstly, the dataset contained many extra data columns such as city, country etc. We had deleted unnecessary columns from it and took only customer ID, product ID, product name, quantity, profit column etc. After that, we had changed the data stored in customer ID and product ID column. The information which were stored in those columns were in complex form. For the convenience of our implementation, we had converted those in numerical form. Finally, 70% data were used for training purpose and 30% data were used for testing purpose.

Table 1: Pre-processed Dataset

	Customer ID	Product ID	Product Name	Category	Sub-Category	Profit	NPQ
0	89	3739	Samsung Convoy 3	Technology	Phones	62.1544	0.215805
1	8873	287	Novimex Executive Leather Armchair, Black	Furniture	Chairs	-288.7650	0.528919
2	3681	3682	Nokia Smart Phone, with Caller ID	Technology	Phones	919.9710	0.593887
3	9576	3653	Motorola Smart Phone, Cordless	Technology	Cordless	-96.5400	0.349550
4	14168	3303	Sharp Wireless Fax, High-Speed	Technology	Wireless Fax	311.5200	0.513758
5	8637	3765	Samsung Smart Phone, with Caller ID	Technology	Phones	763.2750	0.395764
6	16815	286	Novimex Executive Leather Armchair, Adjustable	Furniture	Chairs	564.8400	0.337674
7	10639	756	Chromcraft Conference Table, Fully Assembled	Furniture	Tables	996.4800	0.455724
8	89	151	Sauder Facets Collection Library, Sky Alder Fi..	Furniture	Bookcases	54.7136	0.215405
9	90	197	Globel Push Button Managers Chair, Indigo	Furniture	Chairs	5.4801	0.165334
10	90	1222	Newell 330	Office Supplies	Newell	4.6644	0.260140

Table 2: Dataset of Global superstore

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment
40098	CA-2014-AB10015140-41954	11/11/2014	11/13/2014	First Class	AB-100151402	Aaron Bergman	Consumer
26341	IN-2014-JR162107-41675	2/5/2014	2/7/2014	Second Class	JR-162107	Justin Ritter	Corporate
25330	IN-2014-CR127307-41929	10/17/2014	10/18/2014	First Class	CR-127307	Craig Reiter	Consumer
13524	ES-2014-KM1637548-41667	1/28/2014	1/30/2014	First Class	KM-1637548	Katherine Murray	Home Office
47221	SG-2014-RH9495111-41948	11/5/2014	11/6/2014	Same Day	RH-9495111	Rick Hansen	Consumer
22732	IN-2014-JM156557-41818	6/28/2014	7/1/2014	Second Class	JM-156557	Jim Mitchum	Corporate
30570	IN-2012-TS2134092-41219	11/6/2012	11/8/2012	First Class	TS-2134092	Toby Swindell	Consumer
31192	IN-2013-MB1808592-41378	4/14/2013	4/18/2013	Standard Class	MB-1808592	Mick Brown	Consumer
40099	CA-2014-AB10015140-41954	11/11/2014	11/13/2014	First Class	AB-100151402	Aaron Bergman	Consumer
36258	CA-2012-AB10015140-40974	3/6/2012	3/7/2012	First Class	AB-100151404	Aaron Bergman	Consumer
36259	CA-2012-AB10015140-40974	3/6/2012	3/7/2012	First Class	AB-100151404	Aaron Bergman	Consumer
28879	ID-2013-AJ107801-41383	4/19/2013	4/22/2013	First Class	AJ-107801	Anthony Jacobs	Corporate
45794	SA-2012-MM7260110-41269	12/26/2012	12/28/2012	Second Class	MM-7260110	Magdelene Morse	Consumer
4132	MX-2013-VF2171518-41591	11/13/2013	11/13/2013	Same Day	VF-2171518	Vicky Freymann	Home Office
27704	IN-2014-PF1912027-41796	6/6/2014	6/8/2014	Second Class	PF-1912027	Peter Fuller	Consumer
13779	ES-2015-BP1118545-42216	7/31/2015	8/3/2015	Second Class	BP-1118545	Ben Peterman	Corporate
39519	CA-2012-AB10015140-40958	2/19/2012	2/25/2012	Standard Class	AB-100151402	Aaron Bergman	Consumer
12069	ES-2015-PJ1883564-42255	9/8/2015	9/14/2015	Standard Class	PJ-1883564	Patrick Jones	Corporate
22096	IN-2015-JS156857-42035	1/31/2015	2/1/2015	First Class	JS-156857	Jim Sink	Corporate
49463	TZ-2015-RH9555129-42343	12/5/2015	12/7/2015	Second Class	RH-9555129	Ritsa Hightower	Consumer
46630	PL-2013-AB600103-41494	8/8/2013	8/10/2013	First Class	AB-600103	Ann Blume	Corporate
36260	CA-2012-AB10015140-40974	3/6/2012	3/7/2012	First Class	AB-100151404	Aaron Bergman	Consumer

Table 3: Dataset of Global superstore

Postal Code	City	State	Country	Region	Market	Product ID	Category
73120	Oklahoma City	Oklahoma	United States	Central US	USCA	TEC-PH-5816	Technology
Wollongong	New South Wales	Australia	Oceania	Asia Pacific	FUR-CH-5379	Furniture	Chairs
Brisbane	Queensland	Australia	Oceania	Asia Pacific	TEC-PH-5356	Technology	Phones
Berlin	Berlin	Germany	Western Europe	Europe	TEC-PH-5267	Technology	Cordless
Dakar	Dakar	Senegal	Western Africa	Africa	TEC-CO-6011	Technology	Wireless Fax
Sydney	New South Wales	Australia	Oceania	Asia Pacific	TEC-PH-5842	Technology	Phones
Porirua	Wellington	New Zealand	Oceania	Asia Pacific	FUR-CH-5378	Furniture	Chairs
Hamilton	Waikato	New Zealand	Oceania	Asia Pacific	FUR-TA-3764	Furniture	Tables
73120	Oklahoma City	Oklahoma	United States	Central US	USCA	FUR-BO-5957	Furniture
98103	Seattle	Washington	United States	Western US	USCA	FUR-CH-4421	Furniture
98103	Seattle	Washington	United States	Western US	USCA	OFF-AR-5309	Office Supplies
Kabul	Kabul	Afghanistan	Southern Asia	Asia Pacific	FUR-TA-3420	Furniture	Tables
Jizan	Jizan	Saudi Arabia	Western Asia	Asia Pacific	TEC-PH-3807	Technology	Phones
Toledo	Parana	Brazil	South America	LATAM	FUR-CH-4530	Furniture	Chairs
Mudanjiang	Heilongjiang	China	Eastern Asia	Asia Pacific	OFF-AP-4959	Office Supplies	Appliances
Paris	Ile-de-France	France	Western Europe	Europe	OFF-AP-3575	Office Supplies	Refrigerator
76017	Arlington	Texas	United States	Central US	USCA	OFF-ST-3078	Office Supplies
Prato	Tuscany	Italy	Southern Europe	Europe	OFF-AP-4743	Office Supplies	Stove
Townsville	Queensland	Australia	Oceania	Asia Pacific	TEC-CO-3597	Technology	Fax Machine
Uvinza	Kigoma	Tanzania	Eastern Africa	Africa	OFF-AP-4967	Office Supplies	Stove
Bytom	Silesia	Poland	Eastern Europe	Europe	FUR-TA-4644	Furniture	Tables
98103	Seattle	Washington	United States	Western US	USCA	OFF-ST-3744	Office Supplies

Table 4: Dataset of Global superstore

Sub-Category	Product Name	Sales	Quantity	Discount	Profit	Shipping Cost	Order Priority	Unit Cost with Discount	Unit Cost
Phones	Samsung Convoy 3	221.98	2	0	62.1544	40.77	High	110.99	110.99
Chairs	Novimex Executive Leather Armchair, Black	3709.395	9	0.1	-288.765	923.63	Critical	412.155	457.95
Phones	Nokia Smart Phone, with Caller ID	5175.171	9	0.1	919.971	915.49	Medium	575.019	638.91
Cordless	Motorola Smart Phone, Cordless	2892.51	5	0.1	-96.54	910.16	Medium	578.502	642.78
Wireless Fax	Sharp Wireless Fax, High-Speed	2832.96	8	0	311.52	903.04	Critical	354.12	354.12
Phones	Samsung Smart Phone, with Caller ID	2862.675	5	0.1	763.275	897.35	Critical	572.535	636.15
Chairs	Novimex Executive Leather Armchair, Adjustable	1822.08	4	0	564.84	894.77	Critical	455.52	455.52
Tables	Chromcraft Conference Table, Fully Assembled	5244.84	6	0	996.48	878.38	High	874.14	874.14
Book-cases	Sauder Facets Collection Library, Sky Alder Finish	341.96	2	0	54.7136	25.27	High	170.98	170.98
Chairs	Global Push Button Manager's Chair, Indigo	48.712	1	0.2	5.4801	11.13	High	48.712	60.89
Newell	Newell 330	17.94	3	0	4.6644	4.29	High	5.98	5.98
Tables	Bevis Conference Table, Fully Assembled	4626.15	5	0	647.55	835.57	High	925.23	925.23
Phones	Cisco Smart Phone, with Caller ID	2616.96	4	0	1151.4	832.41	Critical	654.24	654.24
Chairs	Harbour Creations Executive Leather Armchair, Adjustable	2221.8	7	0	622.02	810.25	Critical	317.4	317.4
Appliances	KitchenAid Microwave, White	3701.52	12	0	1036.08	804.54	Critical	308.46	308.46
Refri-gerator	Breville Refrigerator, Red	1869.588	4	0.1	186.948	801.66	Critical	467.397	519.33
Storage	Akro Stacking Bins	12.624	2	0.2	-2.5248	1.97	Low	6.312	7.89
Stove	Hoover Stove, Red	7958.58	14	0	3979.08	778.32	Low	568.47	568.47
Fax Machine	Brother Fax Machine, High-Speed	2565.594	9	0.1	28.404	766.93	Critical	285.066	316.74
Stove	KitchenAid Stove, White	3409.74	6	0	818.28	763.38	High	568.29	568.29
Tables	Hon Computer Table, with Bottom Storage	1977.72	4	0	276.84	759.47	Critical	494.43	494.43
Storage	Carina 42"Hx23 3/4"W Media Storage Unit	242.94	3	0	4.8588	1.28	High	80.98	80.98

As shown in table 2,3 and 4 the parameters we used from the datasets are Product ID, Customer ID, Product Name, Category, Sub-Category, Quantity and Per Unit Cost. While experimenting with the dataset for further use, we divided the data into 2 datasets. We put Product ID, Customer ID and Quantity in a separate dataset to make sparse matrix and to generate a recommendation. We used 3 most popular recommendation algorithms: KNN, FKNN, and SVD.

Features:

Order Date: The date when a particular item is purchased. The date format is DD-MM-YY. E.g.: if someone bought a product on 12th July 2018 then data would be 12-07-18.

Customer ID: A unique id for the customers which is helpful to distinguish customers with similar names. The ID consists of the initial of customer's name along with a unique numerical value. E.g.: Justin Ritter's customer id is JR-162107.

City: The name of the city from where the customer requested the order or want to receive the product. E.g.: Berlin.

Country: The name of the country from where the customer requested the order or want to receive the product. E.g.: Australia.

Product ID: A unique id for a product. The ID consists of the first 3 letter of the category of the product, first 2 letters of the sub category and unique numerical value. E.g.: Product ID TEC-PH-5816 means the product is from the technology category and phones sub-category, FUR-CH-5379 means the product is from the furniture category and chairs sub-category.

Category: A particular section the product falls in. E.g.: Technology, Furniture etc.

Product Name: The name of a product a customer ordered. E.g.: Motorola Smartphone, Nokia Smartphone etc.

Plotting Heatmap from Correlational Coefficients: To understand correlational coefficients easily, we use the heatmap to visualize it. In correlational coefficients, value close to 1 is considered most relevant or recommended product and value close to -0.5 is considered most irrelevant product. In the heatmap, the value of correlation coefficient is replaced color; 1 is yellow and -0.5 is blue. So, more yellowish color means the product is highly recommended.

Heatmap for First 20 Products: Here, this graph shows a correlation between first 10 products from the dataset in figure 7. The color ranges from yellow to blue where yellow is 1.0 and blue is -0.5. The numbers in both X and Y-axis denotes Product ID. Now, let's analyze the first row. First row shows the correlation between 10 products including the product itself. Column 0 is obviously yellow for row 0 since they are the same product. In row 0, column 8 and 9 is respectively closer to yellow which means they are relevant to Product ID 0.

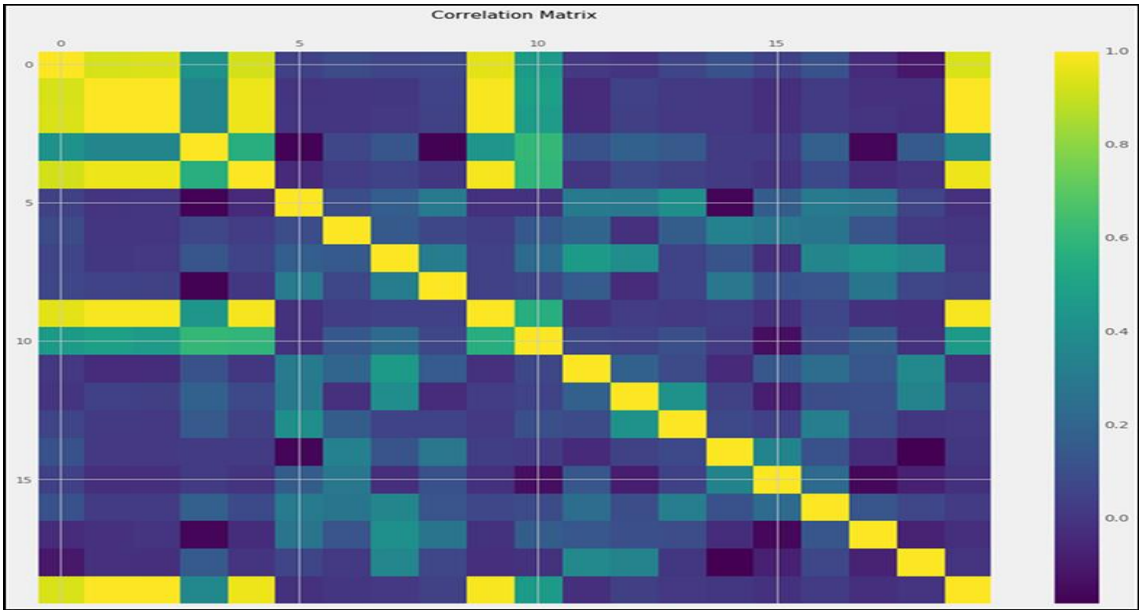

Figure 7: Heatmap for first 20 products.

Chapter 5

Experimental Results

5.1 Result for Recommendation

.

- Creating Customer-Product matrix from dataset.
- Converting Customer-Product matrix to CSR Matrix.
- Computing Cosine Similarity and distances for KNN and FKNN.
- Creating fuzzy set for FKNN for extended similarity computation.
- Decomposing Customer-Product matrix to Product matrix using SVD.
- Getting correlation coefficients using matrix factorization.
- Generating recommended products for KNN, FKNN and SVD.
- Plotting future profit graphs for recommended product pairs using ARIMA model.

Firstly, we used our customer-product matrix as input and converted it into csr matrix for faster computation. Then we used scikit.neighbour to implement knn using the csr matrix by computing cosine similarity and distance between products based on similarities. Similarly, we followed the same steps to implement FKNN over our data and additionally we imported fuzzywuzzy library to create fuzzy sets for input products. Finally, we implemented our SVD recommender engine using truncatedSVD library. After getting all the recommended lists for products we forecasted the profits of products using ARIMA model and to implement that we imported sciemetric.api library. To compare accuracy and errors of the algorithms, we imported Surprise Lib library to calculate Precision, Recall, MAE and RMSE.
First, we used KNN algorithm to train the data. Later, we tried Fuzzy KNN, also known as FKNN, for more improved result. To find distance and similarities between products and rank nearest neighbors, we selected cosine metrics. KNN and FKNN list all the products with distances between other products. As input, name of a product and number of top n recommended products are given. And the output shows top n products with their distances. After using KNN and FKNN, we used SVD algorithm which is also very popular algorithm for a recommendation. The output format is same as KNN and FKNN but here correlation coefficient is shown instead of distance. Correlation coefficient measures the relationship between two products. Values less than 0.2 are considered as relevant products and values close to 0 are highly recommended. The distance value comparison is for all algorithms are shown in Table 3.

Table 5: Recommended Products for KNN & FKNN

Product Name	**KNN**	
Sharp Wireless Fax Machine, High-Speed	Office supplies	Accos Push Pins, Bulk Pack - distance - 0.8903121479721579 Acme Box Cutter, High Speed - distance - 0.9113564542765047 Binney & Smith Pencil Sharpener, Easy-Erase - distance - 0.9164785799522052 Enermax Cards & Envelopes, 8.5 x 11 - distance - 0.9214720619979764
	Furniture	Dania Library with Doors, Metal - distance - 0.9175443755595604 SAFCO Steel Folding Chair, Black - diatance - 0.9227511838010567
	Technology	Memorex Mini Travel Drive 8 GB USB 2.0 Flash Drive - distance- 0.8914595543293653 SanDisk Router, Programmable - distance - 0.9091151454714065 Okidata Card Printer, Red - distance - 0.9281839241589959 Canon Ink, Color - distance - 0.9204384196485731
	FKNN	
	Office supplies	Accos Push Pins, Bulk Pack - distance - 0.8687797052119952 Acme Box Cutter, High Speed - distance - 0.8961930867695982 Binney & Smith Pencil Sharpener, Easy-Erase - distance - 0.9045379876161312 Advantus Thumb Tacks, Assorted Sizes – distance- 0.9240321578564631 Honeywell Quietcare HEPA Air Cleaner - distance - 0.9280612012252809
	Technology	SanDisk Router, Programmable - distance - 0.8965386993306694 Memorex Mini Travel Drive 8 GB USB 2.0 Flash Drive - distance - 0.904918156607308 Canon Ink, Color - distance - 0.9156357590501434
	Furniture	Dania Library with Doors, Metal - distance - 0.9123642786253783 SAFCO Steel Folding Chair, Black - distance - 0.91265517218661

Table 6: Recommended Products for SVD

Product Name	**SVD**	
Sharp Wireless Fax Machine, High-Speed	Office supplies	Accos Push Pins, Bulk Pack - distance - 0.0581728427672062 Acme Box Cutter, High Speed - distance - 0.0929669864623999 Binney & Smith Pencil Sharpener, Easy-Erase - distance - 0.1994801004751738 Advantus Thumb Tacks, Assorted Sizes - distance - 0.2230546885313575 Tenex Shelving, Industrial- distance- 0.2265885106501911
	Furniture	SAFCO Steel Folding Chair, Black - distance - 0.2085121601057755 Dania Library with Doors, Metal - distance- 0.2128398598948649
	Technology	Canon Ink, Color - distance - 0.2151978827699903 Memorex Mini Travel Drive 8 GB USB 2.0 Flash Drive - distance - 0.2168223872387488 SanDisk Router, Programmable - distance - 0.0936065407112067

Since we already made recommendation using KNN, FKNN and SVD, we just modified the recommended products' data by filtering out the products of same sub category to show the cross sell products as in Table 7. Here is the conditional pseudo code

If product_subcategory != recommended_product_subcategory

print recommended_product

Table 7: Sample Cross-sell output for KNN, FKNN and SVD

Product Name	**KNN**
Nokia Smart Phone,with Caller ID	1. Green Bar Message Books, Recycled with a distance of 0.8855893122401853 2. Belkin Keyboard, USB with a distance of 0.8954253455581193 3. Harbour Creations Bag Chairs, Adjustable with distance of 0.9067036192897082 4. Stanley Canvas, Fluorescent with a distance of 0.9123679954819723 5. Accos Staples, Metal with a distance of 0.9146967012296934
	FKNN
	1. Green Bar Message Books, Recycled with a distance of 0.8485413394015521 2. Belkin Keyboard, USB with a distance of 0.8685030206075899 3. Harbour Creations Bag Chairs, Adjustable with a distance of 0.87786131431276 4. Accos Staples, Metal with a distance of 0.8994744006551902 5. StarTech Inkjet, White with a distance of 0.9093383766655856
	SVD
	1. Green Bar Message Books, Recycled with a correlation coefficient of 0.0219231573456607 2. Belkin Keyboard, USB with a correlation coefficient of 0.0460357041719879 3. Harbour Creations Bag Chairs, Adjustable with correlation coefficient of 0.0599982499671869 4. Accos Staples, Metal with a correlation coefficient of 0.0957813168640791 5. StarTech Inkjet, White with a correlation coefficient of 0.102601509379207

Table 8: Sample Up-sell outputs for KNN, FKNN and SVD.

Product name	**KNN**
Nokia Smart Phone, with Caller ID, unit cost 638.91	Apple iPhone 5 -unit cost 649.83
	FKNN
	Apple iPhone 5 –unit cost 649.83
	SVD
	Apple iPhone 5 - unit cost 649.83

For upsell, the result shows the narrowed down list of recommended products with higher price within the same sub category. The result is determined by the unit cost as shown in Table 8. Here is the conditional pseudo-code:

If product_unit_cost < recommended_unit_cost and product_subcategory ==

recommended_product_subcategory:

Print recommended_product

5.2 Result for Prediction

Future Profit Forecast:
As our main objective is to find out how much the profit will be if a product is paired with the recommended products, we used the cross-sell recommendation list for further computation. Despite the data for the individual product is insufficient to generate forecast, we made the program to create top product sub-categories from cross-selling recommendation list by automatically choosing the sub-categories of most recommended products.

Top Sub-Categories from Cross-Sell Recommendation:

Input Nokia Smart Phone, with Caller ID .

Table 9: Recommended product list.

KNN	FKNN	SVD
1. Paper	1. Paper	1. Paper
2. Accessories	2. Accessories	2. Accessories
3. Chairs	3. Chairs	3. Chairs
4. Canvas	4. Fasteners	4. Fasteners
5. Fasteners	5. Ink	5. Ink

So, the results are shown in table 9, top 5 unique sub-categories based on most recommended products generated by each recommendation algorithm.

Plotting Profit Forecast using ARIMA model:

ARIMA model is used to analyze and forecast time-series data. The dataset we are using has large sales data from year 2012 to 2016 and since we are using pair of sub-categories instead of individual products, the data is sufficient to generate profit forecast using ARIMA model. We tried to forecast profit from 2017 to 2024 (8 years in total). First, it generates a line graph of profit forecast of the input product only as in Figure 8 so that a comparison can be made later.

Input: Nokia Smart Phone, with Caller ID.

Output:

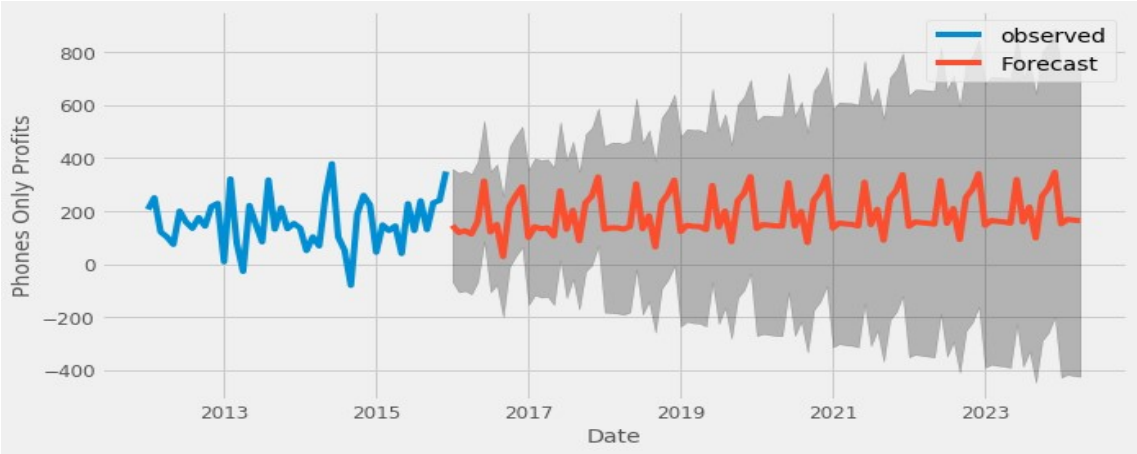

Figure 8: Future Profits for Phones (X=Years, Y=Profits in USD).

The line graph in Figure 9 shows predicted future profits of phones. The blue line denotes the observed data from 2012 to 2016, the red line denotes average profit from 2017 to 2024 and the gray areas behind the red line denotes the maximum-minimum peak of profit margin. After generating the forecast graph the input product's sub-category, pairs of products are being made by adding profits of input product's sub-category and recommended product's subcategory one by one from recommended sub-category list for each algorithm. Later, profit forecast is being generated for each graph.

Profit Forecast Graphs of Recommended Pairs:

Since Paper, Accessories, Chairs, Fasteners are found in top 5 of KNN, FKNN and SVD, so we are going to check profit forecast of these products.

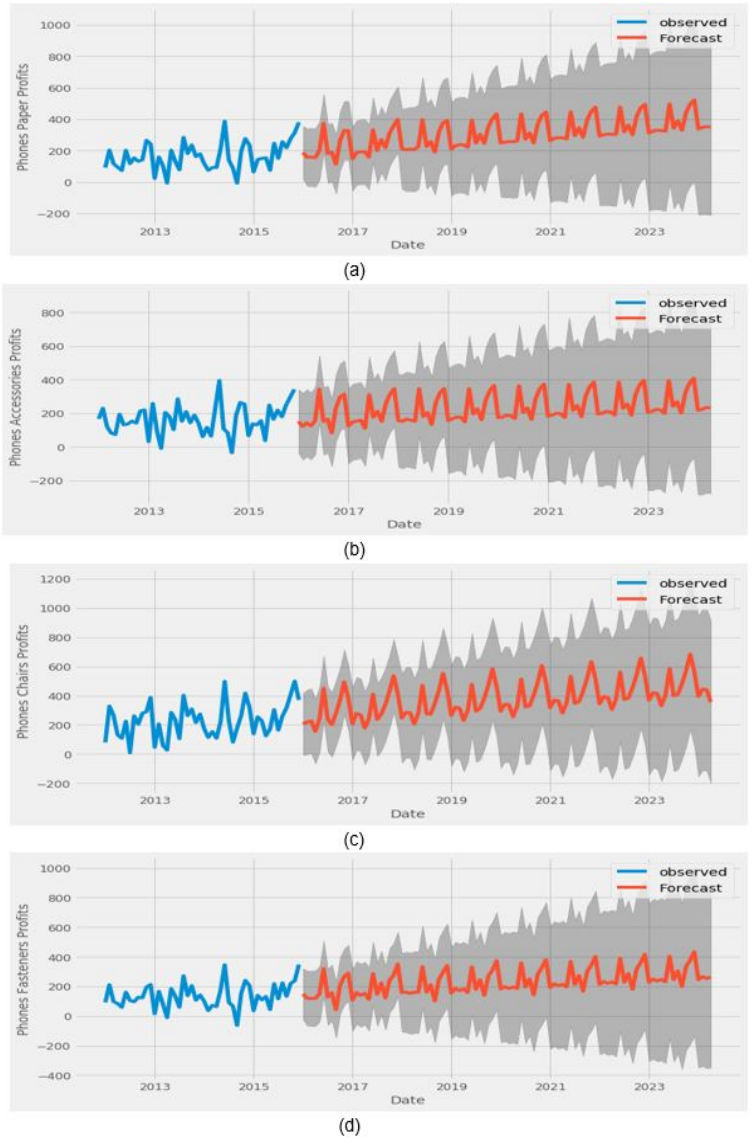

Figure 9: Future Profits for Phones and Papers (X=Years, Y=Profits in USD). (b) Future Profits for Phones and Accessories (X=Years, Y=Profits in USD), (c) Future Profits for Phones and Chairs (X=Years, Y=Profits in USD) (d) Future Profits for Phones and Fasteners.

Chapter 6

Result Analysis

First, we tried to find out the precision, recall, MAE and RMSE of KNN, FKNN and SVD. But before we examine these metrics, we should be acquainted with two terms: recommended and relevant.

Relevant: The items that have rating greater than given threshold. Here the threshold is 0.5 to 1 from rating scale 0 to 1.

Recommended: The items that are being predicted in the recommendation list with highest estimated rating.

Precision: Precision is the fraction value that tells us how many of the items are relevant from the recommendation list. The equation of precision is:

Precision@k=Recommended items that are relevant/Recommended items
From table 10, we get the precision of KNN is 81.95 %, FKNN is 85.30% and SVD is 91.02%. It is clearly visible that SVD has the maximum precision which means more relevant items are generated by the SVD recommendation system.

Recall: Recall is the fraction value that tells us how many of the recommended items are actually relevant to the particular item. The equation of recall is:

Recall@k= Recommended items that are relevant/Relevant items

The recall values we got from all the algorithms are:

Here, the recall value of KNN is 84.46%, FKNN is 95.59% and SVD is 76.95%. In this case, Fuzzy KNN got the maximum recall value which means the recommended products generated by FKNN are more relevant.
Since accuracy is not good enough to measure the quality of a recommendation algorithm, we tried to find error metrics such as RMSE and MAE of the algorithms.

Table 10: Precision, Recall, MAE and RMSE of KNN, FKNN, SVD.

Algorithm	Precision	Recall	MAE	RMSE
KNN	81.95	84.46	0.8353	1.3401
FKNN	85.30	95.59	0.7689	1.1271
SVD	91.02	76.95	0.6981	0.9737

MAE: Mean Absolute Error or MAE estimates the average value of absolute error.
To find the best system by measuring error, we look for the system that has fewer errors. Here, the MAE value of KNN is 0.8353, FKNN is 0.7689 and SVD is 0.6981. Since SVD has a lesser MAE value, it is the best recommendation system in this case.

RMSE: Root Mean Squared Error or RMSE estimates the root average of squared errors of a system.

Here, the RMSE value of KNN is 1.3401, FKNN is 1.1271 and SVD is 0.9737. As SVD still got the lowest value, we can definitely say that SVD is a better recommendation system than KNN and FKNN.

Chapter 7

Conclusion

In this era of modern business and marketing, consumers do not have to search for products rather products run after their targeted consumers. To take control over this highly competitive market, understanding the consumer's buying trend is very important. Through our research we have implemented some methods which can help the producers to understand the pulse of their consumers and subconsciously motivate them to buy their products. At the same time, producers also can get idea about probability of their future cross sell profits. In our research we have used KNN, Fuzzy KNN and SVD methods for product recommendation. We got close results from all of the algorithms but among three SVD gave the best result. However, for predicting the future profit of paired items (Time series analysis) we have used ARIMA model. From the result of time series forecasting we found that the cross selling pairs will increase the profit in future for the sellers. However, our implemented methods bring some features which can make shopping more convenient for consumers. Consumers can get their desired products together under a package by cross selling method and they can get more idea about their desired products without searching with the help of up selling methods. In addition, though today's world is moving to online based shopping our research findings can add a new dimension in this field. Also, traditional super shops/producers can boost their marketing strategy and attract more consumers by implementing our findings.

To conclude, it is said that a healthy market environment lives when both consumers as well as producer/retailers become satisfied with the condition of market and an excellent market environment can change the socio-economic scenario of a country. We hope our research and findings will have significant contribution to set a healthy market environment.

References

1. Salazar, M. T., Harrison, T., Ansell, J. "CRM in the Insurance Industry: An Attempt to Use Survival Analysis in Retention and Cross Selling", Journal of FRONTIERS OF E-BUSINESS RESEARCH, 2004.

2. Varlander, S., Yakhlef, A. "Cross-selling: The power of embodied interactions", Journal of Retailing and Consumer Services, 2008.

3. Johnson, J. S., Friend, S. B. "Contingent cross-selling and up-selling relationships with performance and job satisfaction: A MOA-theoretic examination,", Personal Selling Sales Management, 2014.

4. Kamakura, W.A "Cross-Selling", Journal of Relationship Marketing, pp.41-58, 2008.

5. Schmitz, C." Group influences of selling teams on industrial salespeople's cross-selling behavior", Original empirical research, 2012.

6. Storey, A., Cohen, M. "Exploiting Response Models -Optimizing Cross-Sell and Up-sell Opportunities in Banking.", Emerald Journals, 2002.

7. Wong, W. K., Leung, S. Y., Guo, Z. X., Zeng, X. H., MokElsevier, P. Y. "Intelligent product cross-selling system with radio frequency identification technology for retailing.", Elsevier, 2011.

8. Bernazzani, S. "Cross-Selling and Up selling: The Ultimate Guide", Hub Spot Blog, 2018.

9. Dobreva, K. "A guide to cross-selling up selling techniques", Amasty, 2016.

10. Cohn, C. "A Beginner's Guide To Up selling And Cross-Selling", Forbes, 2015.

11. Charles, C. "What Amazon Can Teach You about Cross-Selling", Predictable Profits, 2014.

12. Harrison, O. "Machine Learning Basics with the K-Nearest Neighbors Algorithm," 2018.

13. Srivastava, T. "Introduction to k-Nearest Neighbors: Simplified (with implementation in Python)," 2014.

14. Maillo, J., Luengo, J., García, S., Herrera, F. and Triguero, I. "Exact fuzzy k-nearest neighbor classification for big datasets," IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp.1-6, 2017.

15. Keller, J. M., Gray, M. R. and Givens, J. A. "A fuzzy K-nearest neighbor algorithm," in IEEE Transactions on Systems, Man, and Cybernetics, pp. 580-585, 1985.

16. Nikoo, M. R., Kerachian, R., Alizadeh, M. R. “A fuzzy KNN-based model for significant wave height prediction in large lakes,” ScienceDirect, 2017.

17. Shang W., Huang H., Zhu H., Lin Y., Wang Z., Qu Y. “An Improved kNN Algorithm – Fuzzy kNN”, Computational Intelligence and Security, 2005.

18. Patel, H., Thakur, G.S. “An Improved Fuzzy K-Nearest Neighbor Algorithm for Imbalanced Data using Adaptive Approach”, IEEE Journal of Research, 2018.

19. Hussan, A. C., Yamur, A. D Lundberg, J. “Time Series Forecasting using ARIMA Model: A case study of mining face drilling rig.”, 2018.

20. Alsharif, M. H., Younes, M. K., Kim, J. “Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation: The Case Study of Seoul, South Korea”, Symmetry, 2019.

21. Weisstein, E. W. ”Singular Value Decomposition”, Math World, 2012.

22. Venkat, R. S. “What is Singular Value Decomposition”, MathWorld, 2013.

23. Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A. “Recommendation systems: Principles, methods and evaluation”, Egyptian Informatics Journal, 2015.

24. Mik, E.C. “New Product Demand Forecasting”, Business Analytics, 2019.

25. Weichbroth, P., Kubiak, B. F., “Cross- And Up-selling Techniques In E-Commerce Activities”, Journal of Internet Banking and Commerce, 2010.

26. Cheng, D., Zhang, S., Deng, Z., Zhu, Y. and Zong, M., “kNN Algorithm with Data-Driven k Value”, 2014.

27. Baker, K., “Singular Value Decomposition Tutorial”, 2013.