Market Basket Analysis Approach to Machine Learning

BY

Md. Abul Hasnat Patwary

ID: 172-15-9807

Md. Tamim Eshan

ID: 171-15-9225

Prazzal Debnath

ID: 172-15-9793

This Report Presented in Partial Fulfillment of the Requirements for The Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Abdus Sattar

Assistant Professor

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

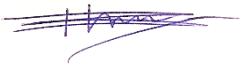
MAY 2021

APPROVAL

This Project/internship titled **"Market Basket Analysis Approach to Machine Learning"**, submitted by Md Abul Hasnat Patwary, ID No:172-15-9807, Md Tamim Eshan, ID No: 171-15-9225, Prazzal Debnath, ID No: 172-15-9793 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on June 2,2021.

BOARD OF EXAMINERS

Chairman

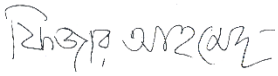


Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Md. Sadekur Rahman
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Fizar Ahmed
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Shamim H Ripon
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

We hereby declare that this thesis has been done by us under the supervision of **Abdus Sattar, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

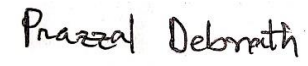
Supervised by:

Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Md. Abul Hasnat Patwary
ID: 172-15-9807
Department of CSE
Daffodil International University

Md. Tamim Eshan
ID: 171-15-9225
Department of CSE
Daffodil International University



Prazzal Debnath
ID: 172-15-9793
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty God for His divine blessing makes us possible to complete the final thesis successfully.

We grateful and wish our profound indebtedness to **Abdus Sattar**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Human-Computer Interaction in Education" to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this thesis.

We would like to express our heartiest gratitude to **Pro. Dr. Akhter Hossain, Professor, and Head of,** Department of CSE, for his kind help to finish our thesis and also to other faculty members and the staff of the CSE department of Daffodil International University.
We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and passion of our parents.

ABSTRACT

Market Basket Analysis is an important aspect of a retail organization's analytical framework for deciding where products should be placed and developing sales promotions for various segments of consumers to increase customer loyalty and, as a result, benefit. The market is the well-known activity of ARM ultimately used for business intelligent decisions. Market Basket Analysis is a data mining technique that can be used in various fields, such as marketing, bioinformatics, education field, nuclear science, etc. Our objective here is that traders can further improve their business. At the same time, they can make more profits and invest more and more. This will be of great benefit. If the business of traders increases, many people will get employment, as well as customers, will be able to make their way of life easier by getting familiar with different things. Here is a discussion on their methods so that traders in Bangladesh can make proper investments in the right place. The frequent item sets are mined from the database using the Apriori algorithm and then the association rules are generated. The project will assist supermarket managers in determining the relationship between the items that their customers purchase.

TABLE OF CONTENTS

CONTENTS PAGE

LIST OF FIGURES

FIGURES **PAGE**

LIST OF TABLES

TABLES **PAGE**

CHAPTER-1

INTRODUCTION

1.1 Introduction

Analysis of affinity is a technique that seeks to co-occur in activities carried out by an Individual or group-specific. It is a methodology for data processing and data mining. It Collects records of a single person or group's everyday activities, such as when They sleep when they work when they eat food, what kind of food they eat, or what kind of food they eat. About customs, etc. This involves being a basket for the market. Market Basket Analysis is a tool of modeling based on the idea that In addition to buying a customer's particular or basic or necessary object, how can you buy a customer's specific or fundamental object? He is interested in purchasing other goods or related products.

In this theory, the buyer tries to extract the relation of the other item from the purchased item. That is how they try to set the items. The main objective of the Market Basket Analysis (MBA) in field marketing, nuclear science, etc., is to provide distributors with data. Understand the buying behavior of the buyer, which can assist the retailer to make the right decision Manufactured. There are various algorithms available for doing an MBA. Current algorithms run Stable data and shifts in data over time are not recorded. But the algorithm proposed is not just mine. However, static data provides a new way to consider data changes. As well recognized to apriori algorithm training or association analytics, Abstract Market Basket Analysis (MBA) is indeed a data mining method that could be used in different fields, like advertising, computational biology, health, nuclear science, education, etc.

It is a modeling technique that is based on a theoretical How to buy a customer's specific or basic or necessary items as well He is interested in buying other items or similar items. In this theory, the buyer Trying to figure out the relationship between another item from the purchased item. Market basket analysis is a data mining method that focuses on discovery purchases. Patterns of customers from associations or co-attendees. Data on a store transaction, for example, when a person checks out items in one Supermarket all go into the transaction database with details about their purchases. Later, many of the huge data are analyzed to determine the type of customer

purchase. Customers also decide which items should be stocked more, such as cross-selling, up-selling, store shelves. Currently, a lot of information is held in databases in different areas, such as retail markets, financial and medical fields, etc. It is not necessary, however, that the entire data is successful for the user. Therefore, extracting useful data from a vast amount of material is very critical. Marketing analysts are indicators of understanding the conduct of customers. No purchase of goods. For performing data mining collectively, there are different methods and formulas available. We all surround ourselves in the world of the Internet with a lot of info. "Where information is available, there are applications for data mining-which are "powered by the application. Attached business basket research “the profit or loss depends on finding the right relationship or link between the products sold with the most common and efficient application of Business Intelligence (BI) when we deal with BI. It is also represented as pattern discovery and exposes the relationship between them so that machine learning and advanced analytics can be effectively combined with modern BI platforms. Discover new trends and aid in the study further. Helps to improve machine learning effectiveness. A data-based machine. It helps to recognize complex patterns automatically and make them intelligent. Information-based decisions. Keywords: data mining, relationships, connections, machine learning, Intelligence for companies. The work of the MBA framework is based entirely on buyer forecasts. The instance recognizes the types of items the buyer purchases, no product received together which product is being purchased with another product means finding the right combination of product, which one product is obtained in certain seasons. Such an approach tends to generate exceptionally valuable data. Assume that a collection of items from a particular category can be purchased by the buyer, but how likely the buyer is to purchase something else, a category with the same item. For example, if the customer purchases raincoats during the rainy season, how likely is it that other raincoat items will be purchased? To boost company profits, such information lets them improve revenue, set up products, and then redesign product packages. Suppose their customers need a prediction device to understand the owner or manager's example of a retail store.

1.2: Motivation:

Thousands of forms of markets exist today. Cell phones, food, movies, etc, for example. They are distinctive channels. For example, online, place, etc. Our emphasis will be on how we can easily draw customers to sell our goods in such a big location. How to get them interested in new products being purchased. They may be interested in purchasing traditional, fancy, aesthetic, and other products at the same time. Our task is to make the company more successful. Traders will come up with new products and something new can be thought of by buyers. Buyers can also dream of something different as well. High-support laws are chosen because they are likely to apply to a significant number of potential transactions. Consumer Basket Analysis is one of the main strategies used to discover connections between goods by large retailers. This works by searching for combinations of things that sometimes occur together in transactions. To put it another way, it enables distributors to describe the connections between the goods that people purchase. Business basket analysis is used by manufacturers so that their consumers may make a buying proposal. It is often used to forecast a client's potential buying option. market basket analysis, the data mining algorithm called the Apriori algorithm is used. The data set just needs to include the basket and the product details to run the MBA. Market Basket Analysis is an unsupervised learning method that needs nothing in the way of feature engineering and a small amount of data cleaning and planning, to put it because of other machine learning techniques. We are encouraged to focus on how companies can make more profit from such a wide market, so many goods, so much competition, etc.

1.3: Problem Definition:

Though a useful and efficient method of marketing data mining, market basket analysis does have a few weaknesses. The first is the type of information required for an efficient basket analysis. To collect meaningful data, it is important to provide a large number of actual transactions, but the quality of the data is undermined if all items do not occur with equal frequency. Therefore, if milk is sold in almost every transaction in our convenience store example, but glue sells only once or twice a month, adding both of them into the same basket analysis would likely yield results that look impressive without being statistically significant-acting on these outcomes does not

necessarily gain profitability. The data-mining program would be able to very confidently state what sells well with glue with just one or two glue clients, but this may only be accurate for the one or two evaluated clients. However, by classifying objects into a taxonomy as defined over the next section, this challenge can be overcome.

Second, the study of consumer baskets may often present findings that are actually due to the effectiveness of previous marketing campaigns. If cola discount coupons have already been placed on the frozen pizza by the convenience store, the fact that cola and frozen pizza sell well together will come as no surprise to them-it does not provide any new details, only show that current marketing strategies are already running. In reality, a real partnership could also be overshadowed by the previous campaign-maybe people would usually prefer buying beer with pizza, but because of the discount, they just buy the cola. The convenience store is losing out on what could be a better promotion in this situation.

1.4: Objective:

The objective of this study is to achieve the following:

1. To take the information from the market and apply machine learning to predict what we need to do in the future.
2. To apply data analysis in our thesis. Where we will have our research process of inspecting, cleansing, transforming and modeling data to discover useful information, informing conclusions, and supporting decision-making.
3. To apply data mining. There will be machine learning, statistics, and database systems There we will try to establish each other's relationship with each event**.**

1.5: Research Question:

To gain usable knowledge, to be able to and to get the right information from the user, and to get it predicted correctly, we need some research questions:

Research Question 1: How do we make a decision where our decision is more likely to be right?

Research Question 2: How buyers create curiosity?

Research Question 3: What effect will it have on buyers?

Research Question 4: What kind of algorithms and what kind of technology can we use to use this model?

1.6: Expected Outcome:

i. When a decision is made through forecasting, it will often succeed. Many problems will be avoided. Traders will benefit most from this. Their money will reduce the risk of loss.

ii. If we add new items and show them that thing, the interest of buyers to buy it will increase. It must be placed in a place where buyers travel a lot so that the thing will be visible to them.

iii. Shoppers can change their lifestyle like convenience with new items.

iv. We will try to create this model using machine learning.

1.7: Layout of the Report:

Our thesis report is organized as follows:

Chapter One includes an introduction to our project, motivation, research questions, and expected outcomes.

Chapter Two includes “Literature Review”, related works, research summary, and challenges.

Chapter Three includes Research Methodology.

Chapter Four includes Experimental Results and Discussion.

Chapter Five includes Conclusion and Future Work.

CHAPTER: 2

LITERATURE REVIEW

2.1: Introduction

There is no similar work or research was done which can detect insects perfectly. First, the works considered the most similar to the work in this thesis are discussed in section 2.2. In Section 2.3, we will give a summary of our related works. In the challenges section, we will discuss how we can increase our accuracy.

2.2 Related Works

Several works relating to our study have been conducted. [1] This paper explains how machine learning can be implemented relatively. Clear approaches for finding consumer buying trends and how they can be turned into actionable methods. Insight irrespective of its size for online or other retailers. They discussed the use of machine learning in different cases to establish connections and the use of market baskets in different cases. [2] this paper says that at present there are a lot of databases in different sectors. For example, medical, bank, etc. But not all the information is needed. To get this necessary information, they said to use machine learning. It will be useful in different businesses They will waste a lot of time with the pay. Association rule mining has been used and algorithms have been used. Several data mining algorithms have been developed and applied to several practical problems. Due to requirements in various applications and data mining limitations, this area is growing. [3] in this paper, they using data cleaning and neural network strategy, the proposed new predictive model for MBA was proposed. They built one, The method of data cleaning helps to improve the consistency of the input dataset and hence the results of the MBA by extracting all Error types from it. MBA model based on artificial neural secondly unsupervised machine learning. Built to network. Using the neural network approach, the current Apriori algorithm is updated to Optimize the effects of the forecast. To the best of our comprehension, this is our first MBA attempt. They are designing an optimized technique for an MBA to predict and analyze the customer's buying behaviors. This has many challenges. Data is

the first challenge. Cleaning, as none of the current approaches even considered the possibility of raw data or noisy data in the history of Via purchases. Second, in terms of seasons and seasons, consumer demands are constantly changing and time. Market basket analysis performance is also entirely dependent on time and seasons, so they have to perform Over and over again. Therefore, a dynamic and automatic MBA system is required. Centered on data cleaning, they implemented new algorithms. [4] In this paper, they present ways that market basket analysis applications can be used to illustrate and integrate common topics common in disconnection, formal logic, set, set in market-basket analysis of topics related to the definition and implementation of a bachelor computer science framework course syllabus. Includes activities, power sets, proof methods, dynamic programming, algorithm analysis, and data structures. Further, market-basket analysis can be divided into phases that allow speech elements and experience focusing on the application of these topics in the classroom allows frequent item set construction and subsequent theoretical and programming assignments to apply them to the second stage of learning the rules. Here they talk about selling different types of bets online and using the app. [5] the tasks in this paper are as follows, Identification of regular transaction products based on support and trust. To build the association rule from the sets of frequent objects. There are some problems he talks about that people now purchase regular products from local supermarkets. Many stores supply their customers with products. The issue facing many retailers is the location of products. They are unaware of the customer's shopping preferences because they do not know which goods in their store should be put together. With the aid of this application shop managers will evaluate the close relationships between the products. Ultimately, this lets them bring items that co-occur together next to each other. It also influences decisions such as which item to store more, cross-selling, up-selling, store shelf arrangement. There is some limitation of his work. The application is a desktop application and will not be available online.

The application input is a file containing integer values representing the list of items, and the integer values are manually mapped. [6] In this paper, they have worked on how to use data mining on market basket analysis. Here they talked about their various applications, they gave examples of them. They wanted to group the customers according to their buying habits. [7] in this paper, they are talking about three-domain in market basket analysis. The development of personalized recommendations is the first domain. Nowadays, this technique is well known. Personalized reviews have also emerged as part of the marketing process since the rise of e-commerce. The

definition essentially consists of recommending goods to consumers based on their expectations. There are two basic ways of getting things done. The first is to suggest items similar to those purchased by the consumer in the past. The second is to check for similar customers and to recommend items purchased by others. For enterprises, both techniques are also used to incorporate cross-selling and upselling strategies. In the field of spatial distribution in chain stores, the second domain where market basket analysis is used is. Owing to the growing number of items that exist today, physical space in stores has begun to be an issue. Increasingly, shops spend money and time trying to figure out which product distribution will lead them to gain more sales. Because of that, knowing in advance which goods are usually bought together, it is possible to adjust the distribution of the store to sell more products and the last domain they are talking about, establishing deals and promotions. Customers-based unique sales may be carried out by actions. For example, if the consumer understands which goods are often bought together, he will create new deals based on them to maximize the sales of those products. They showed the high power of BigML. How machine learning as a service breaks from the conventional approaches used today without sacrificing efficiency or versatility to create data science projects. Unfortunately, we couldn't show the simplicity in this project, Implantation The models of BigML has been in development and have shown that machine learning algorithms can be used to solve problems in the real world and How the use of them will provide businesses with a quality advantage over their rivals. The information they provide is extremely important to a company. [8] in this paper, he has described his purpose. The first one is, to research the underlying notion of data mining technology by Explore the clustering algorithm definition. Another one is, to run the market-basket research program to extract hidden patterns among different supermarket products by the use of a program for data mining called Poly Analyst. He is talking about the problem. The problem is the value of data is not understood by most organizations today. To the advantage of organizations, mining technology. This is possibly due to the introduction in the market of hundreds of organizational tools and applications that confuse. Several corporate individuals. This analysis would therefore be straightforward and direct, focus indirectly on discussing the relevance of data mining to the Enterprise. [9] in this paper, he has stated his intentions. The statements are, A lot of focus has been given to the topic of data mining over the past two decades. Although retailers are involved in this subject due to the absolute usefulness of market basket data, market analysts are interested in the analysis and technological challenges they face when analyzing the data. Every second, a

growing amount of data is generated and this enables experts to look for meaningful consumer purchase associations. On a single shopping trip, consumers make buying choices in many product categories. Interdependencies between. As retailers aim to develop their businesses by applying quantitative analysis to their results, goods have recently faced increased scrutiny. For retailers, it is really important to get to know what their clients are purchasing. Certain items have a greater affinity to be sold together and thus if special deals and promotions are created for these products, the retailer will benefit from this affinity. It is also important for the dealer to cut off goods that do not produce a profit from the assortment. The removal of loss-making, declining, and poor brands will help corporations increase their revenues and redistribute expenses to more profitable brands. This is yet another reason why data mining is seen as a powerful method to regularly check whether too many products are sold, find poor ones, and probably combine them with healthier brands for many companies. For the valuable knowledge they provide, data mining techniques are highly regarded so that the retailer can better support clients and produce a higher profit. He is talking about some problems. The problem is in recent years, it has become very attractive for retailers to examine shopping baskets. Advanced technology has made it possible for them to obtain data about their clients and what they purchase. In market basket research, the advent of electronic point-in transactions expanded the use and application of transactional data. Such knowledge is extremely useful for analyzing purchasing behavior in retail market research. Mining purchase trends enables retailers to better change promotions, shop settings, and serve customers. Any effective company needs to define purchasing rules. For mining useful information on co-purchases and modifying promotion and ads accordingly, transactional data is used. Only an example of an association law found by data scientists is the well-known collection of beer and diapers. The main objective of the study is to see how various items interrelate in a beauty shop collection and how marketing activities can manipulate these relationships. Transactional data mining association guidelines would provide us with useful knowledge about co-occurrences and commodity co-purchases. During a shopping trip, some shoppers can buy a single product out of curiosity or boredom, while others purchase more than one product for efficiency reasons. [10] in their paper, their purpose is customer acquisitions of a sample retail store in corporate society and seeing how buying habits are linked to shelf layout and promotion. This study aims to find the effects of the consumer basket analysis implementation in a retail store. Proper rearrangement of shelves and promotions will lead to an increase in sales volume. It is intended to discover the

pattern of co-occurrence among some items in shopping baskets using the SPSS Modeler development package. To help them determine and execute the correct layout and promotions, the results will be reported to the store manager. Also, the goal is to closely track the shifts in sales volume and to infer the results of the study of the consumer basket.Their problem discussion is, Because of the financial crisis in the last few years and since the pace of wage promotions is around half the inflation rate, there has been a tremendous decrease in the sales volumes of corporate retail stores. Some retail store divisions of corporate society have been shut down, and few main branches still serve government workers and distribute their essential needs.

2.4 Challenges:

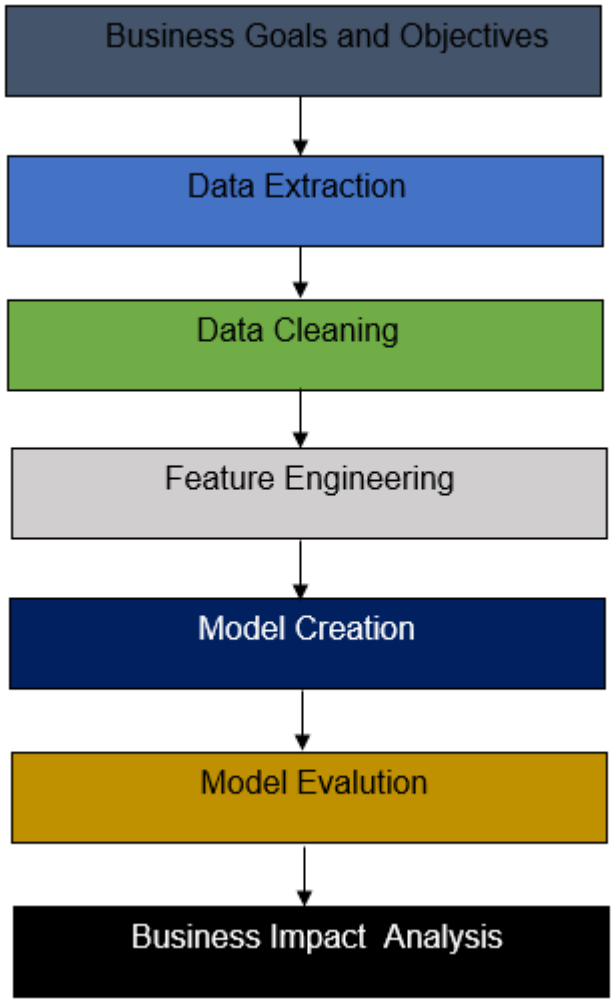
Not much has been done in our country in this regard. Because the environment and the economy depend on it. It requires a lot of big data collection. In this paper, we will try to make it easier for companies, users, and private owners to make it easier for companies to use and at the same time, we can use machine learning to present it more simply.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction:

The process we normally use to carry out a data science project is described in this chapter. Some steps still have to be taken to carry out a data science project. Each phase, with its features and goals, forms part of the total project procedure.



3.1.1Figure Data Science Structure

Every rectangle represents a project phase. On the one hand, measures from data extraction to model evaluation are connected to every general project in information science. There is repetition between them in those steps. On the other hand, depending on the organization and its

specifications, the measures with the stars reflect them. Good numbers are approximations of time costs over total project time.

One may have a priori definition that the development and evaluation of the model is the most time spent in a project. It's absolutely the opposite, though. The process of data transformation occupies much of the project time. Rows are the flux between steps. This is one of the most characteristic features of a data

Project in Science. Flux on conventional projects is linear, there is just one iteration, but with this type of project, iterations have to be worked on. Due to any new situation or outcome, a completed procedure or step may be replicated.

3.2 Business Goals and Objectives:

The aims and goals that have to be accomplished are the foundation of every project. It's really necessary this first move. Both the project itself and the plan will be influenced by decisions and policies agreed upon here. In this stage, the client introduces what he wants to accomplish using machine learning techniques. Then our job is to examine those goals to understand and understand them.

Decide if machine learning algorithms can be used to accomplish them. If they can be done, we decide how the project will be carried out and what outcomes we want to accomplish. Often, businesses approach us hoping that utilizing the problem they have can be solved. Algorithms for machine learning when they're not. Much of the time, there is a general misunderstanding about what machine learning is and its capabilities.

3.3 Data Extraction:

The method of gathering all a company's historical data is data extraction. This data is called raw because it has not already had any treatment. Data extraction is the first step that can be considered part of the process of data transformation. Data collection can also be a difficult task because each client has its method of storing the data. Data is also distributed between different resources and has different formats. Other times, information is poorly ordered or even unstructured. All these factors make the extraction of data a challenging job. Nowadays, several tools suit this form of

problem. Each of them has its features and methodologies, however, even with this support, the data collection process. Using these tools can mean a big task, and the process is not a simple job.

3.4 Data Cleaning:

The method of identifying and deleting corrupt or incorrect information from historical data is data cleaning. This cleaning method is one of the most distinct aspects of a data science project carried out at the university and in the real world. Most of the time, datasets used to learn and practice are already clean, they don't need to be handled. The issue is that that's not the case in the real world. There are a lot of mistakes with datasets. Such mistakes are caused by various causes and the project needs to identify them. Invalid records can mean deterioration by introducing noise or false information to the future model. Cleaning data is often used in the process of eliminating data that is not important or required. Part of the work is to know which information is relevant or can provide value to the algorithm and handle it for each particular case. Data is duplicated in another common cause. Due Databases are from large organizations and often the data is replicated from various sources.

3.5 Feature Engineering:

The process of using domain knowledge of the data to create features that make machine learning algorithms work is feature engineering. In a data science project, this method is fundamental, but it is complicated and costly. Most of the project time is spent on this task due to this high cost. The task consists of identifying features that add value and which ones do not. The approach involved the development, transformation, and deletion of features, and the consistency of the model generated with those features was checked over all the cases. Features used to train a machine learning model influence its performance. The better the characteristics are, the better the output would be. In the project, the quality and quantity of characteristics have a huge influence. The functions are the ones that add value to the model, rather than the hyperparameter configuration of the algorithm. It is worth spending time making new ones,
Features, analyzing them, and transforming knowledge rather than attempting various algorithms.

3.6 Model Creation:

Once the characteristics are developed, the machine learning model is educated. Models are fed using the provided info. The algorithm can be monitored or unmonitored and, depending on the target, the classification or regression task would be the project. To capture the evolving behavior of the data, these models of machine learning need to be periodically retrained. With the client and according to the requirements, this period must be specified of the problem.

3.7 Model Evaluation:

Model assessment is the last step in the normal process. The success of it is the outcome of all the work performed in the process. There are several metrics to measure the efficiency of a model, depending on the type of the problem. There are two forms of assessment, offline and online. Before bringing it into development, the first one analyzes the consistency of a model a priori. A split of 80/20 is the basic way to do it.

The dataset or a cross-validation execution. The second one is to validate the model and evaluate its output in the current data. A/B testing is one famous tactic used. This consists of choosing a subset of instances from the total set and testing the model's results for that subset.

3.8 Business Impact Analysis:

In a data science project, the last step is the impact analysis of the solution. To achieve a monetary advantage, businesses usually prefer to carry out programs. It can either be direct or indirect. On the one side, one used for churn prediction is an example of a model used to achieve a direct monetary advantage. It gives the business a direct income because it avoids the loss of churning customers. An example of a model used to achieve indirect advantages, on the other hand, may be one that groups clients for a posteriori marketing campaign based on their habits. This model does not provide direct income input, but knowledge of consumer habits will contribute to future income. In Data, we can assist our customer to understand what data says about a business, but in the end, the customer is the one who has to implement the corresponding behavior.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction:

In Chapter four, we will discuss, if the production is technically made in any way, customers will be interested in buying the product. We have shown the mathematical term here. Here, we will discuss Association Rule Learning, the Apriori algorithm.

4.2 Association Rule Mining:

The relationship between association rules and IF-THEN can be thought of as an IF-THEN relationship. If a customer purchases item A, the chances of item B being selected by the same customer under the same Transaction ID are determined.

There are two elements of these rules:

Antecedent (IF): This is a type of item or group of items that can be contained in Itemsets or Datasets.

Consequent (THEN): This item comes with an Antecedent or a collection of Antecedents.

However, there is a stumbling block. If you make a rule about an object, there are still about 9999 things to consider when making rules. The Apriori Algorithm comes into play at this stage. So, first, let's look at the math that underpins the Apriori Algorithm. There are three methods for determining association:

- Support
- Confidence

- Lift

Support: It tells you what percentage of transactions include items A and B. Support informs us of the most commonly purchased products or combinations of items.

$$Support = \frac{freq(A,B)}{N}$$

So with this, we can filter out the items that have a low frequency.

Confidence: It tells us how often the items A and B occur together, given the number of times A occurs.

$$Confidence = \frac{freq(A,B)}{freq(A)}$$

When working with the Apriori Algorithm, you usually define these terms as required. But how do you figure out how much anything is worth? There isn't a way to describe these words, to be honest. Assume you've set the support value to 2. This means that you won't consider the item/s for the Apriori algorithm unless and until their frequency is less than 2%. This makes sense because it's pointless to think about things that aren't purchased very much.

Let's say you still have about 5000 things after filtering. For anybody, creating association rules for them is a near-impossible job. This is where the principle of lift enters the picture.

Lift: The intensity of a rule over the occurrence of A and B at random is indicated by the lift. It tells us how strong a rule is.

$$Lift = \frac{Support}{Supp(A)\ x\ Supp(B)}$$

Focus on the denominator, it is the probability of the individual support values of A and B and not together. Lift explains the strength of a rule. More the Lift more is the strength. Let's say for A -> B, the lift value is 4. It means that if you buy A the chances of buying B is 4 times.

4.3 Apriori Algorithm:

The Apriori algorithm generates association rules by using frequent item sets, and it is designed to work with transaction databases. It defines how strongly or weakly two objects are related using these association laws.

Item sets that have a higher level of support than the threshold value or the user-specified minimum level of support are called frequent item sets. It implies that if A and B are frequent item sets together, then A and B should also be frequent item sets separately.

Suppose there are the two transactions: A= {1,2,3,4,5}, and B= {2,3,7}, in these two transactions, 2 and 3 are the frequent item sets.

4.4 Steps for Apriori Algorithm:

Below are the steps for the apriori algorithm:

Step-1: Determine the transactional database's support for item sets and choose the lowest level of support and trust.

Step-2: Take all transaction supports that have a higher support value than the minimum or chosen support value.

Step-3: Find both of these subsets' rules with a higher confidence value than the threshold or minimum confidence.

Step-4: Sort the rules as the decreasing order of lift.

4.5 Apriori Algorithm Working:

We will understand the apriori algorithm using an example and mathematical calculation:

Example: Suppose we have the following dataset that has various transactions, and from this dataset, we need to find the frequent item sets and generate the association rules using the Apriori algorithm:

ID	Itemset
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B,C

Table 4.5.1 Example Dataset

Here, Minimum Support=2, Minimum Confidence=50%

Step1: Calculating C1 and L1:

In the first step, we will create a table that contains the support count (The frequency of each item set individually in the dataset) of each item set in the given dataset. This table is called the Candidate set or C1.

Item set	Support Count
A	6
B	7
C	5
D	2
E	1

Table 4.5.2 Support Count

We'll now remove all the items with a higher support count than the Minimum Support (2). It will provide us with the table for the L1 frequently used item set. All of the item sets, except E, have a greater or equivalent support count than the minimum support, so E will be omitted.

Item set	Support Count
A	6
B	7
C	5
D	2

Table 4.5.3 Minimum Support Count

Step 2: Candidate Generate C2 and L2:

With the aid of L1, we will generate C2 in this step. In C2, we'll make the pair of L1 item sets in the form of subsets. We'll get the help count from the main transaction table of datasets after we've created the subsets, which is how many times these pairs have occurred together in the given dataset. As a result, we'll get the following table for C2:

Item set	Support_Count
(A,B)	4
(A,C)	4
(A,D)	1
(B,C)	4
(B,D)	2
(C,D)	0

Table 4.5.4 Minimum Support Count

The C2 Support count must be compared to the minimum support count once more, and the itemset with the lowest support count will be removed from table C2. For L2, we'll get the table below.

Itemset	Support Count
(A,B,C)	2
(B,C,D)	1
(A,C,D)	0
(A,D,B)	0

Table 4.5.5 Minimum Support Count

We'll now make the L3 table. As seen in the C3 table above, there is only one item set combination with a support count equal to the minimum support count. As a result, the L3 will only have one combination, namely (A, B, C).

Step-4: Finding the association rules for the subsets:

To produce the association rules, we'll start by creating a new table with all of the possible rules from the A, B.C combination. We'll use the formula sup (A B)/A to measure the Trust for all of

the laws. We will exclude the rules that have less confidence than the minimum threshold after measuring the confidence value for all rules (50 percent).

Consider the below table:

Rule	Support	Confidence
A ^B → C	2	Sup{(A ^B) ^C}/sup(A ^B)= 2/4=0.5=50%
B^C → A	2	Sup{(B^C) ^A}/sup(B ^C)= 2/4=0.5=50%
A^C → B	2	Sup{(A ^C) ^B}/sup(A ^C)= 2/4=0.5=50%
C→ A ^B	2	Sup{(C^(A ^B)}/sup(C)= 2/5=0.4=40%
A→ B^C	2	Sup{(A^(B ^C)}/sup(A)= 2/6=0.33=33.33%
B→ B^C	2	Sup{(B^(B ^C)}/sup(B)= 2/7=0.28=28%

Table 4.5.6 Count Confidence

As the given threshold or minimum confidence is 50%, so the first three rules A ^B → C, B^C → A, and A^C → B can be considered as the strong association rules for the given problem.

4.6: Advantages of Apriori Algorithms:

1.This is easy to understand the algorithm

2.The join and prune steps of the algorithm can be easily implemented on large datasets.

3.The algorithm is exhaustive, so it finds all the rules with the specified support and confidence

4.7: Disadvantages of Apriori Algorithms:

1.The apriori algorithm works slow compared to other algorithms.

2.The overall performance can be reduced as it scans the database multiple times.

3. Requires many database scans.

4.8: Python Implementation of Apriori Algorithm:

Now we'll look at how the Apriori Algorithm works in practice. To put this into practice, we have a problem with a retailer who wants to find the connection between his store's products so that he can give his customers a "Buy this, Get that" deal.

The merchant has a dataset of information that includes a list of his customer's transactions. Each row in the dataset shows the goods that customers have purchased or the transactions that they have made. We'll take the following steps to solve this issue:

1.Data Pre-processing

2.Training the Apriori model on the dataset

3.Visualizing the results

4.9: Data Collection Procedure:

Now since data collection from our supermarket is not possible . So, we collected data from (https://www.kaggle.com/heeraldedhia/groceries-dataset?fbclid=IwAR26uph_fBND1iFula1Gl20O1hBf2Hfa0DaTPfeJYLQvfEyN3BvfwsyMJ-g).

The dataset has 38765 rows of the purchase orders of people from the grocery stores. These orders can be analysed and association rules can be generated using Market Basket Analysis by algorithms like Apriori Algorithm.

4.10: Data Format and Statistical Analysis:

Member number	Date	item Description
1808	21-07-2015	tropical fruit
2552	05-01-2015	whole milk
2300	19-09-2015	pip fruit
1187	12-12-2015	other vegetables
3037	01-02-2015	whole milk
4941	14-02-2015	rolls/buns
4501	08-05-2015	other vegetables
3803	23-12-2015	pot plants
2762	20-03-2015	whole milk
4119	12-02-2015	tropical fruit
1340	24-02-2015	citrus fruit
2193	14-04-2015	Beef
1997	21-07-2015	frankfurter
4546	03-09-2015	chicken
4736	21-07-2015	butter
1959	30-03-2015	fruit/vegetable juice
1974	03-05-2015	packaged fruit/vegetables
2421	02-09-2015	Chocolate
1513	03-08-2015	specialty bar

Table4.10.1 Sample of Dataset

Data	Unique Number
Date	728
Member Number	3898

Table 4.10.2 Data Statistics

4.11: Data Pre-Processing:

After data collection, we need to preprocess it again. Now, here is some missing data. There is a function called fillna to handle the missing data. Fillna function, it's in the library Panda as a function. Here, is the data len 14963.Then we encoded this data. We have a library called mlxtend

where we encoded by importing Transaction Encoder. Then we reshaping this data through NumPy array so that we tread the data.

4.12: Training the Apriori Model on the dataset:

Now, we will implement Apriori Algorithm. To train the model, we have to follow some parameters:

transactions: A list of transactions.

min_support= To set the minimum support float value. Here we have used 0.001.

min_confidence: To set the minimum confidence value. Here we have taken 0.001. It can be changed as per the business problem.

min_lift= To set the minimum lift value.

min_length= It takes the minimum number of products for the association.

max_length = It takes the maximum number of products for the association.

4.13: Data Visualization:

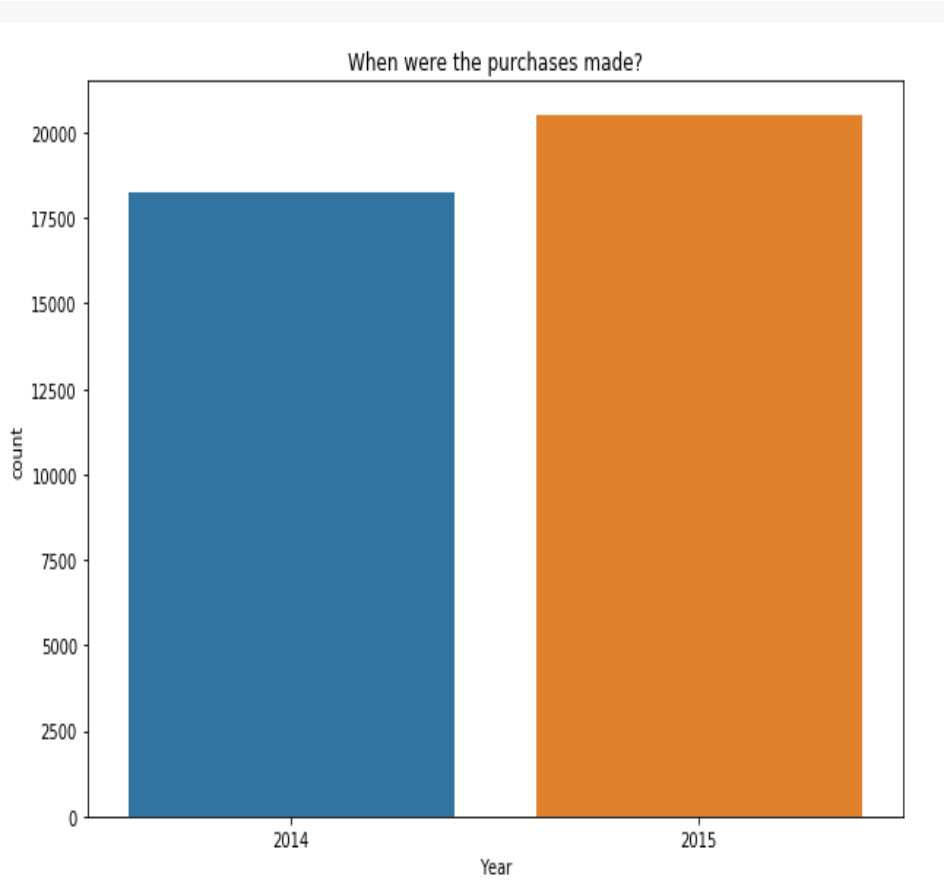


Figure 4.13.1 Purchases Made

Here is the account for the year. We can see that more sales were made in 2015 than in 2014.This means that by analyzing the data of 2014, the owner has been able to make more profit by using them in 2015.

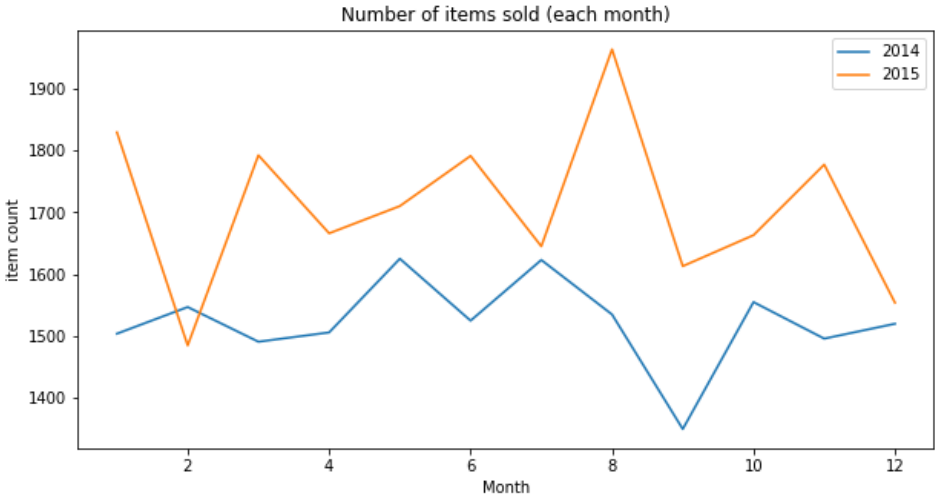


Figure 4.13.2 Number of items sold

Here, see how much is being sold in a month. We can see that the number 8 month saw low sales in 2014 but much higher sales in 2015.

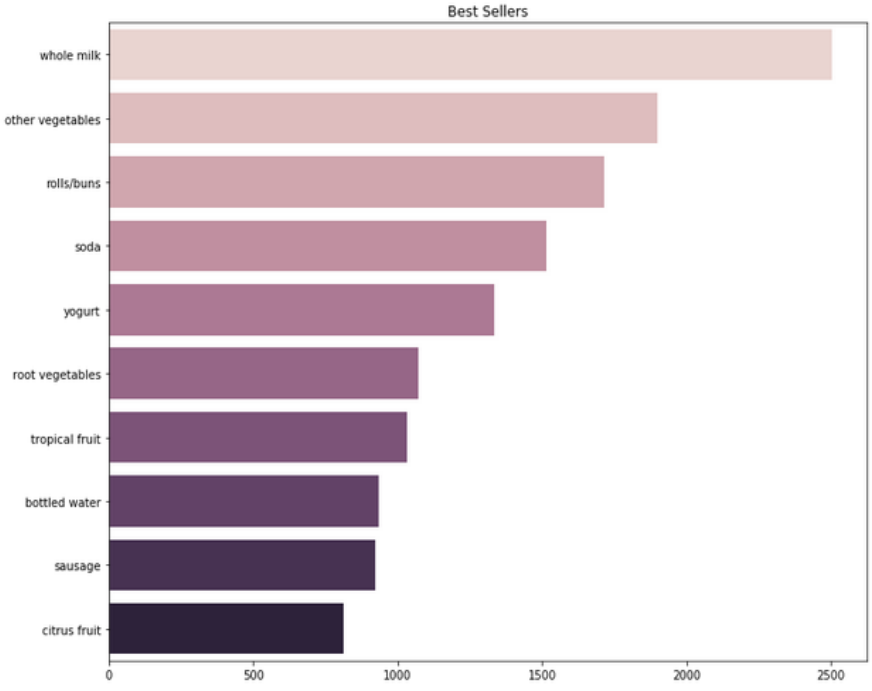


Figure 4.13.3 Best Sellers

We analyze various data here to see which items are selling the most. We can see here that customer demand is higher on whole milk and this is the most sold.Then vegatables are sold more than roll then other items.

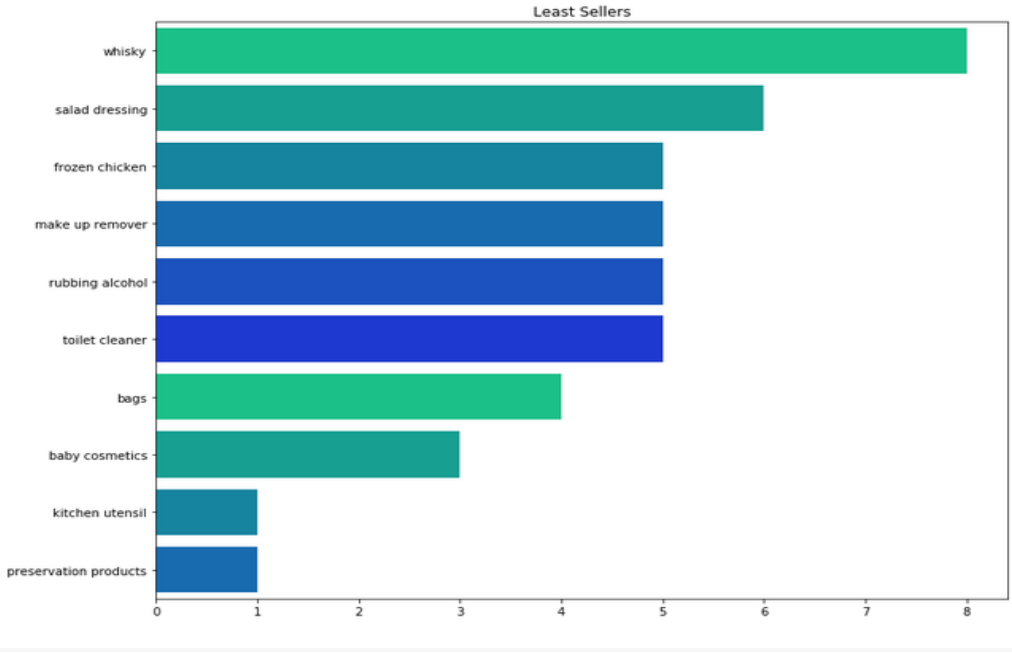


Figure 4.13.4 Least Sellers

Here we see the list of latest item sales . we can see that whisky is the most sold item here, follwed by salads and other items.

Figure 4.13.5 Mostly Sell of Items

It is word frequencies of our project. Word frequencies means, lists of words in a language grouped by frequency of occurrence within a text corpus, either by levels or as a ranked list, for the purpose of vocabulary acquisition. It show that, any item is being sold more, which is more likely to seek customers, which is looking for a customer more than those who are looking for more.

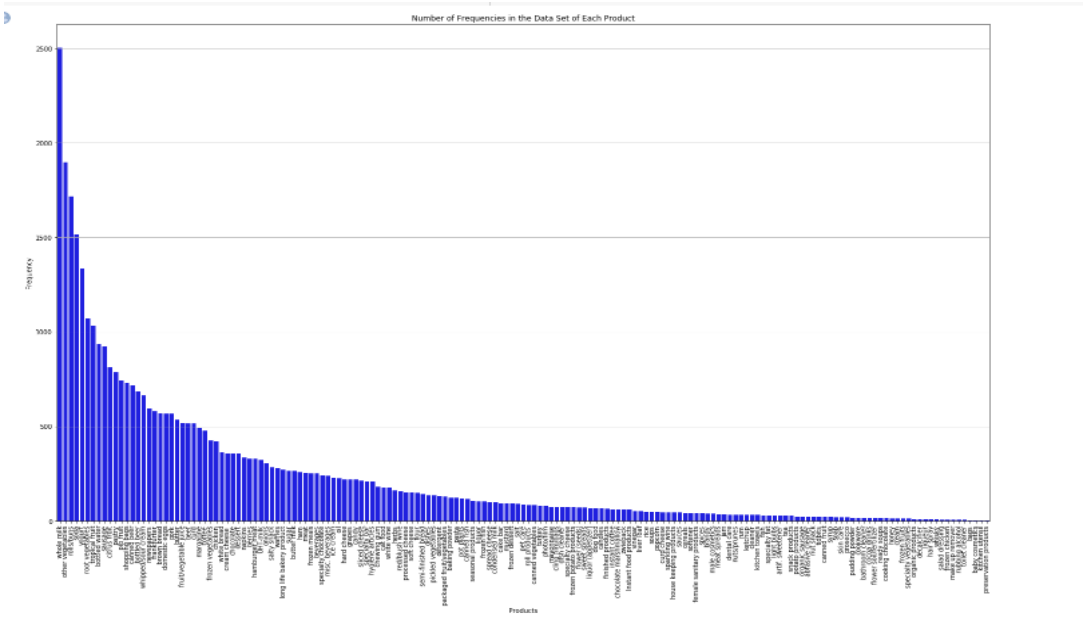


Figure 4.13.6 Selling frequency of each item

Here, the selling frequency of each item is shown. We can see that the sale of whole milk has increased. People are buying it more, the demand for it is higher and then it is selling more.

4.14: Output:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
1234	(yogurt, sausage)	(whole milk)	0.005748	0.157923	0.001470	0.255814	1.619866	0.000563	1.131541
1209	(rolls/buns, sausage)	(whole milk)	0.005347	0.157923	0.001136	0.212500	1.345594	0.000292	1.069304
1228	(soda, sausage)	(whole milk)	0.005948	0.157923	0.001069	0.179775	1.138374	0.000130	1.026642
1108	(semi-finished bread)	(whole milk)	0.009490	0.157923	0.001671	0.176056	1.114825	0.000172	1.022008
1222	(yogurt, rolls/buns)	(whole milk)	0.007819	0.157923	0.001337	0.170940	1.082428	0.000102	1.015701
...	...	...	...	...	...	...	...	...	...
958	(whole milk)	(pasta)	0.157923	0.008087	0.001069	0.006771	0.837316	-0.000208	0.998675
1229	(whole milk)	(soda, sausage)	0.157923	0.005948	0.001069	0.006771	1.138374	0.000130	1.000829
1021	(whole milk)	(pot plants)	0.157923	0.007819	0.001002	0.006348	0.811821	-0.000232	0.998519
982	(whole milk)	(pickled vegetables)	0.157923	0.008955	0.001002	0.006348	0.708829	-0.000412	0.997376
1217	(whole milk)	(soda, rolls/buns)	0.157923	0.008087	0.001002	0.006348	0.784984	-0.000275	0.998250

Figure 4.13.7 Data Frame Output

From the above output, we can analyze each rule. The first rules, which is (yogurt, sausage)->whole milk, status that the yogurt, sausage, whole milk is bought frequency by the most of computer. The support for this rule is 0.001470 and the confidence is 25%. if a customer buys yogurt and sausage. It is 25% chance to buys whole milk. We can check all these things in other rules also.

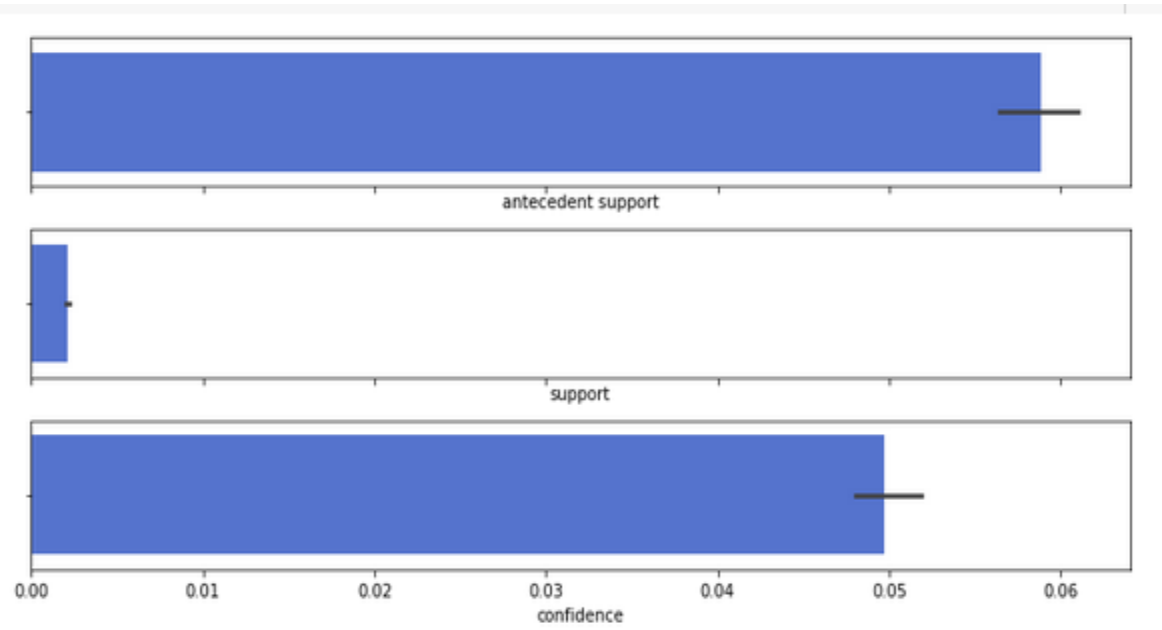


Figure 4.13.8 Box Graph of Support, Confidence, Antecedent Support

Now, we can know which of these products has more confidence and support through the graph of the box. Here length is 0.9. we can see that now the value is 0.00 to 001 then support is low but confidence and antecedent support is high.

CHAPTER 5
CONCLUSION AND FUTURE WORK

Our main purpose is that the owner can analyze the information related to his business. Then accordingly he is able to make more profit than before. We have shown here apriori algorithm how the merchants sort their items, the buyer is interested to buy the product and can get more profit. It's not just the owners who benefit. Here, when the owner benefit, they will introduce buyers to more new items. This will allow buyers to get acquainted quality of life by purchasing product related to their daily life.

Our goal is for big businesses to be more profitable and to be interested in buying other new products. So that customers can get acquainted with other new products. They can improve the quality of life. Here are some things to keep in mind. That is information. Nowadays, people are getting acquainted with new products and new demands day by day. At the same time, the owner can take the customer's review, then they will understand what they like and what they are not doing. Again, what needs to be done and what does not need to be done. It should also be kept in mind that when you have a department store in an area, you have to take care of the economy, tradition, holidays, etc. in that area. It must be made in a good residential area and the price of the product should be commensurate with the quality of the product so that they are interested in buying it. Now we have to pay attention to the tradition because the food habits of each place are the same. Their needs are different. Their sleeping time, eating time, etc. are different in different places. For example, if you want to do business in America, you have to do the bread business, you can't do the rice business there. Because they are not accustomed to eating rice. If you look in Japan, they eat noodles. You also have to trade in noodles. You can't trade in pulses there. If you want to trade in rice and pulses, you have to come to Bangladesh. In the same way, people from different countries have different sleeping times and different working hours. You also have to pay close attention to the religious culture. If you live in a Muslim society, you can trade beef but not pork. It hurts their feelings. You will not be able to trade beef in Hindu society in the same way. You have to take care of the traditional ceremonies of the people of every society in the same way. You have to pay attention to the ethnic characteristics of the society. For example, the people of Bangladesh eat hilsa fish, panta rice on the occasion of New Year. If you hoard other materials in

this materialism you will not benefit there. You have to look at the discount again. Because the discount of the conventional common thing does not make such a difference in the amount of product sold. For example, the common man eats coke with burgers. Now if you reduce the price of cocoa, there will be some profit for selling the product as they will buy this product. But if you see people with burgers like to eat sandwiches but can't take or are not interested in taking them because of the price, you can keep discounts to increase your sandwich sold. Then your sandwich sold will increase. Again, if you wish, you can increase product sales by offering or discounting products that are not sold. Again, you can increase product sales by adding its offer to any product. For example, Maggie Noodles are not being sold in your store. But if you see that Iggy Noodles are being sold, then you can offer that if someone buys 5 packets of Iggy Noodles, he will get a Maggie Noodles for free. This way you can increase product sales. Owners need these things in mind.

REFERENCES

[1] D. Raich, B. Ganguly, and M. Tota, "Machine Learning for Market Basket Analysis through," *IOSR Journal of Engineering (IOSRJEN)*, pp. 22-23, 2019.

[2] M. Kaur and S. Kang, "Market Basket Analysis: Identify the changing trends of market data," in *International Conference on Computational Modeling and Security (CMS 2016)*, Sangrur 148001, 2016.

[3] R. Gangurde, D. B. Kumar and D. S. D. Gore, "Optimized Predictive Model using Artificial Neural for Market Basket Analysis," Research Gate, Pune, Maharashtra, India, 2017.

[4] M. R. Wick and P. J. Wagner, "Using market basket analysis to integrate and motivate topics in discrete structures," in *ACM SIGCSE Bulletin*, Eau Claire, 2006.

[5] S. Mainali, "MARKET BASKET ANALYSIS," GitHub, Kirtipur, 2016.

[6] M. A. Ula¸s, "MARKET BASKET ANALYSIS FOR DATA MINING," Academia.edu, Istanbul, 2001.

[7] G. R. Grau, "Market Basket Analysis in Retail," UPCommons. Global access to UPC knowledge, Barcelona, 2017.

[8] K. A. B. A. KADIR, "CLUSTERING ALGORITHM FOR MARKET-BASKET ANALYSIS: THE UNDERLYING CONCEPT OF DATA MINING TECHNOLOGY," University Putra Malaysia Institutional Repository, Serdang, 2003.

[9] V. Gancheva, "Market Basket Analysis of Beauty Products," SEMANTIC SCHOLAR, Rotterdam, 2013.

[10] F. Arasteh and F. Arbab, "Studying Changes in Corporative Society Retail Store Sales as a Result of Shelves' Rearrangement and Promotions Based on Market Basket Analysis," Digitala Vetenskapliga Arkivet, Tehran, 2016.

Plagiarism Checked by Abdus Sattar, Assistant Professor, Department of CSE	02-05-2021

Market Basket Analysis Approach to Machine Learning

ORIGINALITY REPORT

20%	17%	3%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.javatpoint.com Internet Source	7%
2	www.edureka.co Internet Source	2%
3	Submitted to University of Wales Institute, Cardiff Student Paper	2%
4	Submitted to Southampton Solent University Student Paper	1%
5	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
6	Submitted to Daffodil International University Student Paper	1%
7	medium.com Internet Source	1%
8	Submitted to Heriot-Watt University Student Paper	1%
9	Submitted to ABES Engineering College Student Paper	1%