A Project Report On

MARKET BASKET ANALYSIS USING DEEP LEARNING TECHNIQUES

Submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

Submitted By

MOGILI VEERA LAKSHMI	**20P31A1231**
AKULA V S S D PHANI HASWANTH	**20P31A1201**
UPPALAPATI GAYATRI NANDINI	**20P31A1211**
MATCHA VICTOR RAJ KUMAR	**20P31A1230**

Under the esteemed supervision of

Dr. Ch V Raghavendran M.Tech.,Ph.D

Professor

DEPARTMENT OF INFORMATION TECHNOLOGY

ADITYA COLLEGE OF ENGINEERING & TECHNOLOGY (A)

Permanently Affiliated to JNTUK, Kakinada * Approved by AICTE New Delhi

Accredited by NBA, Accredited by NAAC (A+) with 3.4 CGPA

Aditya Nagar, ADB Road, Surampalem, Kakinada District, Andhra Pradesh.

2020-2024

ADITYA COLLEGE OF ENGINEERING & TECHNOLOGY (A)

(An Autonomous Institution)

Permanently Affiliated to JNTUK, Kakinada * Approved by AICTE New Delhi
Accredited by NBA, Accredited by NAAC (A+) with 3.4 CGPA
Aditya Nagar, ADB Road, Surampalem, Kakinada District, Andhra Pradesh

DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the project work entitled "**Market Basket Analysis Using Deep Learning Techniques**", is a bonafide work carried out by **MOGILI VEERA LAKSHMI (20P31A1231), AKULA V S S D PHANI HASWANTH (20P31A1201), UPPALAPATI GAYATRI NANDINI (20P31A1211), MATCHA VICTOR RAJ KUMAR (20P31A1230)** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology** from Aditya College of Engineering & Technology(A) during the academic year 2020-2024.

Project Guide
Dr. Ch V Raghavendran M.Tech.,Ph.D
Professor

Head of The Department
Mr. R V V N Bheema Rao M.Tech.,(Ph.D)
Associate Professor

EXTERNAL EXAMINER

DECLARATION

We hereby declare that this project entitled "Market Basket Analysis Using Deep Learning Techniques", has been undertaken by us and this work has been submitted to **Aditya College of Engineering & Technology(A)** affiliated to JNTUK, Kakinada, in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology**.

We further declare that this project work has not been submitted in full or part for the award of any degree of this or in any other educational institutions.

PROJECT ASSOCIATES

MOGILI VEERA LAKSHMI	**20P31A1231**
AKULA V S S D PHANI HASWANTH	**20P31A1201**
UPPALAPATI GAYATRI NANDINI	**20P31A1211**
MATCHA VICTOR RAJ KUMAR	**20P31A1230**

ACKNOWLEDGEMENT

It is with immense pleasure that we would like to express our indebted gratitude to our Project Supervisor, **Dr. Ch V Raghavendran** **M.Tech.,ph.D.** who has guided us a lot and encouraged us in every step of the project work, his valuable moral support and guidance throughout the project helped us to a great extent.

We wish to express our sincere thanks to the Head of the Department **Mr. R V V N Bheema Rao** **M.Tech.,(Ph.D)** for his valuable guidance given to us throughout the period of the project work and throughout the program.

We feel elated to thank **Dr. Ch V Raghavendran** **M.Tech.,ph.D** Dean – Academics of Aditya College of Engineering & Technology for his cooperation in completion of our project and throughout the program.

We feel elated to thank **Dr. D Kishore** **Ph.D** Dean – Evaluation of Aditya College of Engineering & Technology for his cooperation in completion of our project and throughout the program.

We feel elated to thank **Dr. Dola Sanjay S** **Ph.D** Principal of Aditya College of Engineering & Technology for his cooperation in completion of our project and throughout the program.

We wish to express our sincere thanks to all **faculty members, lab programmers** for their valuable assistance throughout the period of the project.

We avail this opportunity to express our deep sense and heart full thanks to the Management of **Aditya College of Engineering & Technology (A)** for providing a great support for us in completing our project and also throughout the program.

PROJECT ASSOCIATES

MOGILI VEERA LAKSHMI	**20P31A1231**
AKULA V S S D PHANI HASWANTH	**20P31A1201**
UPPALAPATI GAYATRI NANDINI	**20P31A1211**
MATCHA VICTOR RAJ KUMAR	**20P31A1230**

Aditya College of Engineering & Technology (A)
(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Institute Vision & Mission

Vision

To induce higher planes of learning by imparting technical education with

- International standards
- Applied research
- Creative Ability
- Values based instruction and to emerge as a premiere institute

Mission

Achieving academic excellence by providing globally acceptable technical education by forecasting technology through

- Innovative research and development
- Industry institute interaction
- Empowered manpower

Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALEM

Aditya College of Engineering & Technology (A)
(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Vision

To be a department with high repute and focused on quality education

Mission

- To Provide an environment for the development of professionals with knowledge and skills
- To promote innovative learning
- To promote innovative ideas towards society
- To foster trainings with institutional collaborations
- To involve in the development of software applications for societal needs

Head of the Department

Head of the Department
Dept.of IT
Aditya College of Engineering & Technology
SURAMPALEM 533 437

Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALEM

Aditya College of Engineering & Technology (A)

(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Program Educational Objectives

Program educational objectives are broad statements that describe the career and professional accomplishments that the program is preparing graduates to achieve.

PEO-1:

Graduates will be skilled in Mathematics, Science & modern engineering tools to solve real life problems.

PEO-2:

Excel in the IT industry with the attained knowledge and skills or pursue higher studies to acquire emerging technologies and become an entrepreneur.

PEO-3:

Accomplish a successful career and nurture as a responsible professional with ethics and human values.

Head of the Department

Head of the Department
Dept.of IT
Aditya College of Engineering & Technology
SURAMPALEM 533 437

Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALEM

Aditya College of Engineering & Technology (A)
(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Program Specific Outcomes

PSO-1:

Apply mathematical foundations, algorithmic and latest computing tools and techniques to design computer-based systems to solve engineering problems.

PSO-2:

Apply knowledge of engineering and develop software-based applications for research and development in the areas of relevance under realistic constraints.

PSO-3:

Apply standard practices and strategies in software project development using open-ended programming environments to deliver a quality product.

Head of the Department

Head of the Department
Dept.of IT
Aditya College of Engineering & Technology
SURAMPALEM 533 437

Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALEM

Aditya College of Engineering & Technology (A)

(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Program Outcomes

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem Analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design / Development of Solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. Conduct Investigations of Complex Problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The Engineer and Society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and Sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and Team Work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. Project Management and Finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. Life-Long Learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Head of the Department

Head of the Department
Dept.of IT
Aditya College of Engineering & Technology
SURAMPALEM 533 437

Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALEM

ABSTRACT

Market Basket Analysis is used for many applications like online marketing, recommendation engines, information security, etc. Over the past few years, it has been one of the hot topics among research groups as its widely used e-commerce site to recommend related products or arrangements of layouts on the basis of frequently purchased items in supermarkets and fixing consumer index price as per consumer's demands.

Market basket analysis finds out customers purchasing patterns by discovering important associations among the products which they place in their shopping baskets. It not only assists in decision making process but also increases sales in many business organizations.

Apriori, FP growth and CNN Bi LSTM algorithms are for mining frequent itemsets. For these algorithms predefined minimum support is needed to satisfy for identifying the frequent itemsets. But when the minimum support is low, a huge number of candidate sets will be generated which requires large computation.

In this project, an approach has been proposed to avoid this large computation by reducing the items of dataset with top selling products. The results show that if top selling items are used, it is possible to get almost same frequent itemsets and association rules within a short time comparing with that outputs which are derived by computing all the items.

INDEX

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

CHAPTER 1: INTRODUCTION

1.1 Introduction

Nowadays Machine Learning is helping the Retail Industry in many different ways. You can imagine that from forecasting the performance of sales to identify the buyers, there are many applications of machine learning (ML) in the retail industry. “Market Basket Analysis” is one of the best applications of machine learning in the retail industry. By analyzing the past buying behavior of customers, we can find out which are the products that are bought frequently together by the customers.

Overview

The highly technological era that we live in has made it possible for companies to gather enormous quantities of data. Data mining is becoming more and more common for many businesses worldwide. The large amount of data that is being gathered on a daily basis captures useful information across different aspects of every business. The collection of data on a highly disaggregate level is seen as a raw material for extracting knowledge. While some facts can be revealed directly from disaggregate data, often we are interested to find hidden rules and patterns. Non-trivial insights can be generated through data mining. Data mining contains of various statistical analyses that reveal unknown aspects of the data. Mining tools have been found useful in many businesses for uncovering significant information and hence, providing managers with solutions for complicated problems.

Data mining is commonly seen as a single step of a whole process called Knowledge Discovery in Databases (KDD). According to Fayyad et.al, ‘KDD is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.’ (Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, 1996) Data mining is a technique that encompasses a huge variety of statistical and computational techniques such as: association-rule mining, neural network analysis, clustering, classification, summarizing data and of course the traditional regression analyses.

Data mining gained popularity especially in the last two decades when advances in computing power provided us with the possibility to mine voluminous data. Extracting knowledge and hidden information from data using a whole set of techniques found its applications in various contexts. Knowledge discovery is widely used in marketing to identify and analyse customer groups and predict future behaviour. Data mining is an effective way to provide better service to customers and adjust offers according to their needs and motivations.

Business use of data mining

Companies nowadays are rich in vast amounts of data but poor in information extracted from that data. Big data is seen as a valuable resource and although the concept of data mining is still new and developing, companies in a variety of industries are relying on it for making strategic decisions. Facts that otherwise may go unnoticed can be now revealed by the techniques that sift through stored information. When applying mining tools and techniques we seek to find useful relationships, patterns and anomalies that can help managers make better business decisions.

Data mining tools perform analyses that are very valuable for business strategies, scientific research and getting to know your customers better. Managerial insights are no longer the only factor trusted when it comes to decision-making.

Data driven decisions can lead to better firm performance. Data-based implications are gaining popularity while the gut instinct of managers is remaining in the background. Analyzing data not only improves firm performance but gives us accurate insights on different aspects of the business. Data mining is widely used in marketing for spotting sales trends, developing better marketing campaigns and finding the root cause of specific problems like customer defection or fraudulent transactions, for example. It is also used for prediction of behaviour: which customers are most likely to leave us (customer churns) or what are the things that an individual will be most interested to see in a website

Research problem description

In the recent years analysing shopping baskets has become quite appealing to retailers. Advanced technology made it possible for them to gather information on their customers and what they buy. The introduction of electronic point-in sale increased the use and application of transactional data in market basket analysis. In retail business analysing such information is highly useful for understanding buying behaviour.

Mining purchasing patterns allows retailers to adjust promotions, store settings and serve customers better. Identifying buying rules is crucial for every successful business. Transactional data is used for mining useful information on co-purchases and adjusting promotion and advertising accordingly. The well-known set of beer and diapers is just an example of an association rule found by data scientists.

The main objective of the project is to see how different products in a gifts shop assortment interrelate and how to exploit these relations by marketing activities. Mining association rules from transactional data will provide us with valuable information about cooccurrences and co-purchases of products. Some shoppers may purchase a single product during a shopping trip, out of curiosity or boredom, while others buy more than one product for efficiency reasons.

Motivation for the study

The main point of interest for retailers is to understand dependencies among purchases. Consumers buy various combinations of products on a single shopping trip, but choice scenarios do not seem to be random to market analysts. These multi category decisions result in the formation of consumers' "shopping baskets" which comprise the collection of categories that consumers purchase on a specific shopping trip. (Puneet Manchanda, Asim Ansari and Sunil Gupta,1999).

What is Market Basket Analysis?

Frequent itemset mining leads to the discovery of associations and correlations between items in huge transactional or relational datasets. With vast amounts of data continuously being collected and stored, many industries are becoming interested in mining such kinds of patterns from their databases. The disclosure of "Correlation Relationships" among huge amounts of transaction records can help in many decision- making processes such as the design of catalogs, cross-marketing, and behavior customer shopping Analysis.

A popular example of frequent itemset mining is Market Basket Analysis. This process identifies customer buying habits by finding associations between the different items that customers place in their "shopping baskets" as you can see in the following fig. The discovery of this kind of association will be helpful for retailers or marketers to develop marketing strategies by gaining insight into which items are frequently bought together by customers.

For example, if customers are buying milk, how probably are they to also buy bread (and which kind of bread) on the same trip to the supermarket? This information may lead to increase sales by helping retailers to do selective marketing and plan their ledge space.

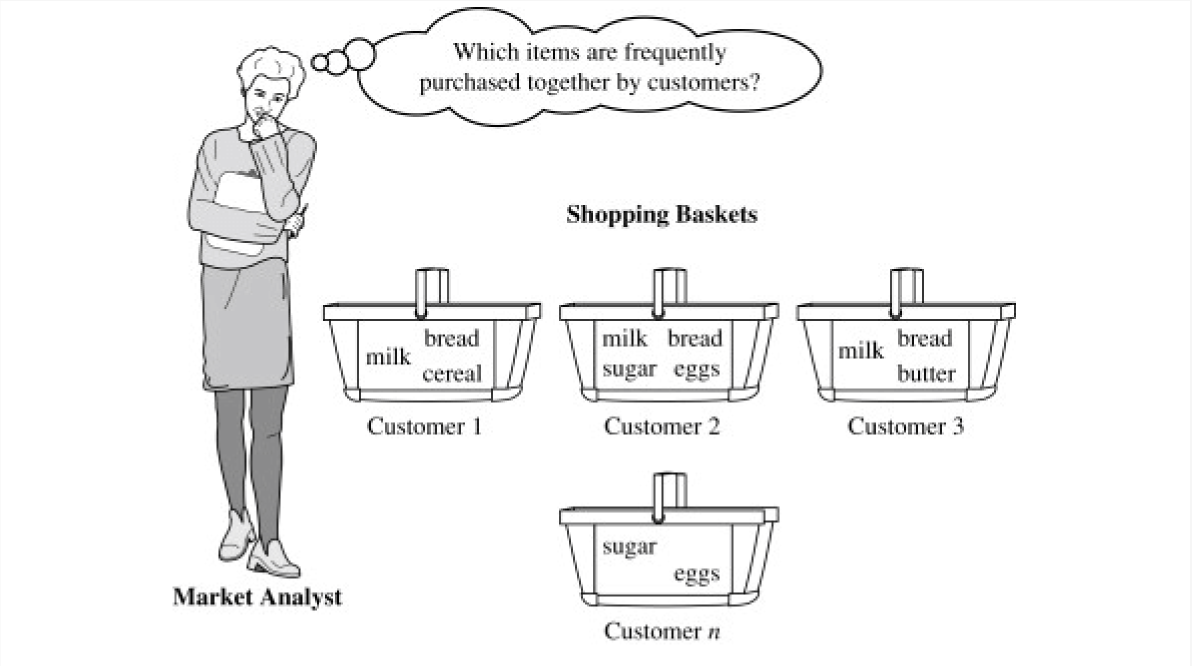


Fig 1.1.1 Market Basket Analysis

Suppose just think of the universe as the set of items available at the store, then each item has a Boolean variable that represents the presence or absence of that item. Now each basket can then be represented by a Boolean vector of values that are assigned to these variables. The Boolean vectors can be analyzed of buying patterns that reflect items that are frequently associated or bought together. Such patterns will be represented in the form of association rules.

What is Association Rule for Market basket Analysis?

Let I = {I1, I2,…, Im} be an itemset. Let D, the data, be a set of database transactions where each transaction T is a nonempty itemset such that **T ⊆ I**. Each transaction is associated with an identifier, called a TID(or Tid). Let A be a set of items(itemset). T is the Transaction which is said to

contain A if **A ⊆ T**. An **Association Rule** is an implication of the form **A ⇒ B**, where **A ⊂ I, B ⊂ I**, and **A ∩B = φ**.

The rule **A ⇒ B** holds in the data set(transactions) D with supports, where ‘s’ is the percentage of transactions in D that contain **A ∪ B** (that is the union of set A and set B, or, both A and B). This is taken as the probability, **P(A ∪ B)**. Rule **A ⇒** B has confidence c in the transaction set D, where c is the percentage of transactions in D containing **A** that also contains **B**. This is taken to be the conditional probability, like P(B|A). That is,

support(A⇒ B) =P(A ∪ B) confidence(A⇒ B) =P(B|A)

Rules that satisfy both a minimum support threshold (called min sup) and a minimum confidence threshold (called min conf) are called "**Strong**".

Confidence***(A⇒ B) = P(B|A) =***
support***(A ∪ B) /***support***(A) =***
support count***(A ∪ B) /*** support count***(A)***

Generally, Association Rule Mining can be viewed in a two-step process:-

1. Find all Frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a pre-established minimum support count, min sup.

2. Generate Association Rules from the Frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence.

Association Rule Mining is primarily used when you want to identify an association between different items in a set, and then find frequent patterns in a transactional database, relational database management system (RDBMS)

Applications of Market Basket Analysis

- **Retail** - Amazon.com
- **Telecom** - bundle TV and Internet packages apart from other discounted online services to reduce churn
- **Medicine** - determine comorbid conditions and symptom analysis

1.2 Literature Survey

Background of the study

Data mining has taken an important part of marketing literature for the last several decades. Market basket analysis is one of the oldest areas in the field of data mining and is the best example for mining association rules.

Various algorithms for Association Rule Mining (ARM) and Clustering have been developed by researchers to help users achieve their objectives. Rakesh Agrawal and Usama Fayyad are one of the pioneers in data mining. They account for a number of developed algorithms and procedures.

According to Shapiro, rule generating procedures can be divided into procedures that find quantitative rules and procedures that find qualitative rules. (Rakesh Agrawal, Ramakrishnan Srikant) elaborate on the concept of mining quantitative rules in large relational tables. Quantitative rules are defined in terms of the type of attributes contained in these relational tables. Attributes can be either quantitative (age, income, etc.) or categorical (certain type of a product, make of a car). Boolean attributes are such attributes that can take on one of two options (True or False, 1 or 0). They are considered a special case of categorical attributes. The authors call this mining problem the Quantitative Association Rules problem. An example of a generated quantitative rule is :

If ((Age: [30…39]) + (Married: Yes)) → (Number of cars = 2) The example

combines variables that have quantitative and boolean attributes.

(S. Prakash, R.M.S. Parvathi, 2011) propose a qualitative approach for mining quantitative association rules. The nature of the proposed approach is qualitative because the method converts numerical attributes to binary attributes.

However, finding qualitative rules is of main interest in this analysis. These rules are most commonly represented as decision trees, patterns or dependency tables. (Gregory PiatetskyShapiro, William Frawley, 1991) The type of attributes used for mining qualitative rules is categorical.

(Rakesh Agrawal, Tomasz Imielinski, Arun Swami, 1993) is one of the first published papers on association rules that proposes a rule mining algorithm that discovers qualitative rules with no restriction for boolean attributes. The authors test the effectiveness of the algorithm by applying it to data obtained from a large retailing company. Association rules found application in many research areas such as: market basket analysis, recommendation systems, intrusion detection etc.

In marketing literature market basket analysis has been classified into two models: explanatory and exploratory. First, exploratory models will be thoroughly explained in this paper as they are of higher relevance for the research and after that an explanation of explanatory models will be given. The main idea behind exploratory models is the discovering of purchase patterns from POS (point-of-sale) data. Exploratory approaches do not include information on consumer demographics or marketing mix variables. (Katrin Dippold, Harald Hruschka, 2010) Methods like association rules (Rakesh Agrawal, Sirkant Ramakrishnan, 1994) or collaborative filtering (Andreas Mild, Thomas Reutterer, 2003) summarise a vast amount of data into a fewer meaningful rules or measures. Such methods are quite useful for discovering unknown relationships between the items in the data.

Moreover, these methods are computationally simple and can be used for undirected data mining. However, exploratory approaches are not appropriate for forecasting and finding the cause-roots of complex problems. They are just used to uncover distinguished cross-category interdependencies based on some frequency patterns for items or product categories purchased together. A typical application of these exploratory approaches is identifying product category relationships by simple association measures. Pairwise associations are used to compare entities in pairs and judge which entity is prefered or has greater amount of some quantitative property. (Julander, 1992) compares the percentage of shoppers buying a certain product and the percentage of all total sales generated by this product.

By making such comparisons, one can easily find out the leading products and what is their share of sales. Examining which the leading products are for consumers is extremely important since a large number of shoppers come into contact with these specific product types every day. As the departments with leading products generate much in-store traffic, it is crucial to use this information for placing other specific products nearby. The paper by Julander also shows how combinatory analysis can be used to study the patterns of cross-buying between certain brands or product groups: for instance, what is the percentage of shoppers that buy products A+C, but not B or what is the percentage of shoppers that buy only A. It also deals with the probabilities that shoppers will purchase from one, two or more departments in a single visit in the store.

Another significant stream of research in the field of exploratory analysis is the process of generating association rules. Substantial amount of algorithms for mining patterns from market basket data have been proposed. From the co-operative work of Rakesh Agrawal and Ramakrishnan Srikant they present two new algorithms for discovering large itemsets in databases, namely Apriori and AprioriTid. These two algorithms are similar with regard to the function that is used to determine the candidate itemsets, but the difference is that the AprioriTID does not use the database for counting support after the first pass (first iteration) while Apriori makes multiple passes over the database. The results from the study show that these two new algorithms perform much better than the previously known AIS (R. Agrawal, T. Imielinski, and A. Swami, 1993) and SETM (M. Houtsma and A. Swami, 1993) algorithms. Since the introduction of the Apriori algorithm, it has been considered the most useful and fast algorithm for finding frequent itemsets. Many improvements have been made on the Apriori algorithm in order to increase its efficiency and effectiveness. (M.J.Zaki, M.Ogihara, S. Parthasarathy, 1996). There are few algorithms developed that are not based on the Apriori,but they still address the issue of speed of Apriori. The following papers (Eu-Hong (Sam) Han, George Karypis, Vipin Kumar, 1999) , (Jong Soo Park, Ming-Syan Chen, Philip S. Yu) propose new algorithms which are not based on the Apriori, but all of them are being compared to Apriori in terms of execution time.

(Robert J. Hilderman, Colin L. Carter, Howard J. Hamilton, and Nick Cercone) develop a framework for knowledge discovery from market basket data. Combining Apriori and AOG (D.W. Cheung, A.W. Fu, and J. Han., 1994) algorithms in the methodology, the purpose of the paper is not only to explain how to discover customer purchase patterns, but to find out customer profiles by dividing customers into distinct classes. The authors provide an extensive explanation of the share-confidence framework. Results show that it can give better feedback than the support- confidence framework.

Another use of market basket data is found in the finite mixture model in the paper by (Rick L. Andrews , Imran S. Currim, 2002). The idea of the model is to identify segments of households that have identical behaviour across product categories. The authors use both marketing variables and scanner panel data to answer the research questions. The study shows that household demographic variables are found to be more strongly correlated to price sensitivity compared to results in previous studies.The research divides customers into heavy users and lighter users. Heavy user households are found to be less price sensitive, visiting the store less often, in most cases high income customers.While, on the other hand, lighter users are mainly students or people that visit the store very often and are very price sensitive. The results show that households that have identical behaviour across product categories tend to be lighter users than households that behave independently. Also households with identical behaviours are said to be more price sensitive,less sensitive to store advertising, also showing weaker loyalty in terms of brand names. The topic on distribution of consumer brand preferences is adressed in the paper by (Gary J. Russel, Wagner A. Kamakura, 1997) using long-run market basket data. The authors show how brand preference segmentation can be discovered without the availability of marketing mix data. A number of simplifying assumptions need to be made in order to permit these cross-category preferences to be estimated. However, using knowledge on marketing mix activity gives the researcher greater flexibility to employ more complex techniques in the analysis than simply using scanner data.

Exploratory models are very useful for uncovering cross-category relations, but not for finding their causes. While the main task of exploratory market basket analysis is to reveal and present hidden relationships between product categories, explanatory models aim at explaining effects. Datasets for such models consist of market basket data, customer attributes and marketing mix variables. The purpose of explanatory models is to identify and quantify cross-category choice effects of marketing variables, such as price, promotion and other marketing features. (Andreas Mild, Thomas Reutterer, 2003) Most of the explanatory models rely greatly on regression analysis, logit,probit and multivariate logistic model.

Mining transactional data along with household data gives retailers and managers space for customised target marketing actions. Analysing past purchases makes it possible for supermarkets to price goods intelligently while still serving heterogeneous consumers. (Nanda Kumar and Ram Rao, 2006). For researchers scanner data is seen as a mean to discover the effects of marketing actions on consumer behaviour. Using the shopping basket as a unit of analysis instead of single articles can provide retailers with consumer-oriented information.

Consumer purchase behaviour is a well-studied area in the marketing literature. The topic of price sensitivity and ellasticity is also well-studied through applied data mining techniques. Customers are commonly divided into large-basket shoppers and small-basket shoppers. Large-basket shoppers have higher expected basket attractiveness in EDLP1 stores, while small-basket shoppers would rather go for HILO2 format of a store. (David R.

Bell and Yasemin Boztu˘g, 2007). In this case with a beauty store, consumers tend to be small rather than large-basket shoppers.

Market basket data combined with household panel data is commonly used by researchers to investigate brand choice and price elasticities (Nanda Kumar and Ram Rao, 2006). Marketing researchers aim to go beyond the trivial correlation approach by finding out the source of cross-category dependence in shopping basket data. Explanatory models are used in this case when the purpose is to explain and predict certain effects. Data sets for such models consist of marketing mix variables and customer attributes. Logit and probit models are commonly used for estimating cross- category effects and predicting brand choice (Gary J Russell, Ann Petersen, 2000).

(Katrin Dippold, Harald Hruschka, 2010) use multivariate logit model to meausre dependencies and sales promotion effects across different categories in a retail assortment and how these effects influence purchase probabilities. As most approaches identify association rules across categories, this multivariate binomial logit model allows for examining main and interaction effects between categories which provides beneficial information on consumer behaviour in terms of predicting the effects of promotion.

Moreover, sensitivity to marketing mix variables is a very common consumer trait, which has been very well studied with the availability of scanner data and household observable variables. There is a strong relationship between household demographic variables and price sensitivity. (Andrew Ainslie, Peter E. Rossi, 1998) measure the covariance of observed and unobserved heterogeneity in marketing mix sensitivity across various categories. Household variables as well as shopping behaviour variables play an important role in explaining price sensitivity.

A common practice for researchers when using explanatory models is to investigate a limited number of cross-category effects. (Gary J Russell, Ann Petersen, 2000) examine brand choice process in four paper goods categories. Brand choice among categories can be easily calculated with a conditional probability formula*, but as the number of categories increases, the level of complexity jumps exponentially. Expanding this general approach to a multivariate logistic model by adding household data gives us the possibility to explore more thoroughly consumer purchase behaviour within a specific store. The authors propose a market basket model based on the idea that choice in one category has impact on choices in all other categories.

Not only because of computational simplicity, but many studies limit included categories to those that are most commonly purchased. However, there has been quite some controversy that results on cross-category objects can be biased because of the small subset of retail assortment that is used in explanatory analysis.Taking into account fewer number of categories can lead to under or overestimation of the values of interaction effects so that some values can even take opposite incorrect signs. Although a research by (Siddhartha Chib, P. B. Seetharaman and Andrei Strijnev) confirms that there is a bias when using a small subset of categories, no such proof is found that there are extreme switches to positive or negative signs of coefficients. However, techniques for mining association rules can easily cope with very large number of categories (or items).

There are some drawbacks and areas of controversy with the exploratory analysis as well. Despite the usefulness of discovering meaningful cross-category interdependencies, the managerial value of exploratory models is somewhat limited. It provides only limited number of recommendations regarding decision-making since there are no apriori assumptions about ‘response’ and ‘effect’ and no marketing variables are incorporated into the analysis. Neglecting both consumer hererogeneity and marketing mix effects may also lead to biases.

(Yasemin Boztuğ , Thomas Reutterer, 2006) propose a model that link both explanatory and exploratory approaches in an attempt to overcome limitations from both approaches. The proposed models employs data compression first and then estimates crosscategory purchase effects in order to reduce the complexity of the model and to select only meaningful categories that are relevant to a specific segment of households. This two-stage procedure that combines feature from both exploratory and explanatory models can be used as a guideline for selecting categories to be included for estimating cross-category effects.

In the book by (Michael J.A.Berry, Gordon Linoff, 1997), the authors suggest an approach of including all kinds of items in the categories. More frequent items do not need to be aggregated at all, while less frequent items need to be rolled up to a higher level of the taxonomy. The term taxonomy refers to a classification of products in a hierarchical fashion. All the single items of a store assortment are on the lowest level of the taxonomy. Based on some shared characteristics, items can be grouped into a category that climbs up the taxonomy. For example, there are five different aromas of a cream soap. They can be all grouped into a category ‘Cream Soap’, which is a subcategory of ‘Soaps’.

Transaction-level data that reflects individual purchases is used in the standard rule mining procedures. However, a lot of models have been proposed for analysis of market basket data at the aggregate level. Data is most commonly aggregated by measures of time so that the base unit is no longer individual transaction, but daily sales in a store for example. It is also possible to roll up transaction-level data by more than one attributes. Here comes the problem of multi-dimensionality discussed in the paper by (Svetlozar Nestorov, Nenad Jukić, 2003). Information on several dimensions – product, location, customer and calendar exists for each transaction. The usual single dimension question – What items are frequently bought together in a transaction? – is now extended to – What products are boughts together in a particular region in a particular month?. When multiple dimensions are involved some associations might be hidden so a new model that captures these dimensions is proposed by the authors. The concept of extended association rules has several advantages in terms of the generated rules: they are easy to explain, providing more accurate predictions for certain variables and the number of discovered rules is likely to be much less for the same threshold support.

Significant amount of papers also contribute to the filed by comparing different mining techniques. Such an example is a recent paper by (A. M. Khattak, A. M. Khan, Sungyoung Lee and Young-Koo Lee, 2010). The authors make comparative analysis of two data mining techniques : ARM (association-rule mining) and Clustering. They use transaction data from a supermarket (Sales Day) to extract important information. Apriori algorithm is used for association rule mining. Its main objective is to find associated products and place them close to each other so that they can benefit from increased sales. When it comes to classification, Clustering is a very preferred technique. The authors apply K-means clustering to classify different classes of products sold together, customers based on their behaviour and purchasing power. The main advantage behind the clustering technique is that in this case there is data available on the customers' profile like age, purchasing power, also customer traffic. Extracting and analysing information from it gives retailer the advantage of improving their business by adopting and implementing new strategies to facilitate customers and maximise sales.

However, a lot of attention has been paid to the problem of generating too many association rules. The problem is addressed in a paper by (Szymon Jaroszewicz, Dan A. Simovici). Hundreds or thousands of association rules can be generated when the minimum support is low. This is why a measure for judging the interestingness of a rule is proposed by the authors. They present an algorithm that computes the interestingness of itemsets with respect to Baysean networks. Interestingness of an itemset is said to be 'the absolute difference between its support estimated from the data and from the Baysean network'.

Given the quantitative nature of the field of data mining, most of the literature on that topic proposes different algorithms and techniques for optimised mining and generation of association rules.

1.3 Problem Statement

- Market Basket Analysis (MBA) faces challenges due to the increasing transactional data in retail, leading to the need for Deep Learning (DL) methods
- Existing methods like Eclat and FP-Growth are used for finding frequent item sets, but there is a gap in combining Association Rules (AR) with DL methods for classification and prediction.
- The need to classify rules and predict the next basket item using DL on transactional datasets is highlighted as a research gap

1.3.1 Disadvantages of Existing system

- Collecting information of customers purchases based on the paper work and the excel file is hard to calculate Market Basket Analysis.
- More Time Taken to measure customer analysis
- User must know the probability statics.

1.4 Objectives of the research

- Implement Deep Learning methods for Market Basket Analysis on transactional datasets.
- Combine Association Rules with Deep Learning for classification and prediction in retail datasets
- Address the gap in using Association Rules as a feature selection with DL methods for analyzing customer purchase behavior
- Develop methods to handle large-scale transactional datasets efficiently, ensuring that the proposed model can be applied to real-world scenarios with millions of transactions.

- Investigate the potential of combining traditional machine learning techniques with deep learning methodologies to create hybrid models that leverage the strengths of both approaches for improved market basket analysis.

1.5 Databases Description

Introduction

A dataset is a collection of data. Most commonly a dataset corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the dataset in question. The data set lists values for each of the variables such as the height or weight of an object for each member in the dataset. A data set is organized into some type of data structure. In a database, for example, a data set might contain a collection of business data (names, salaries, contact information, sales figures, and so forth). The database itself can be considered a data set, as can bodies of data within it related to a particular type of information, such as sales data for a particular corporate department.

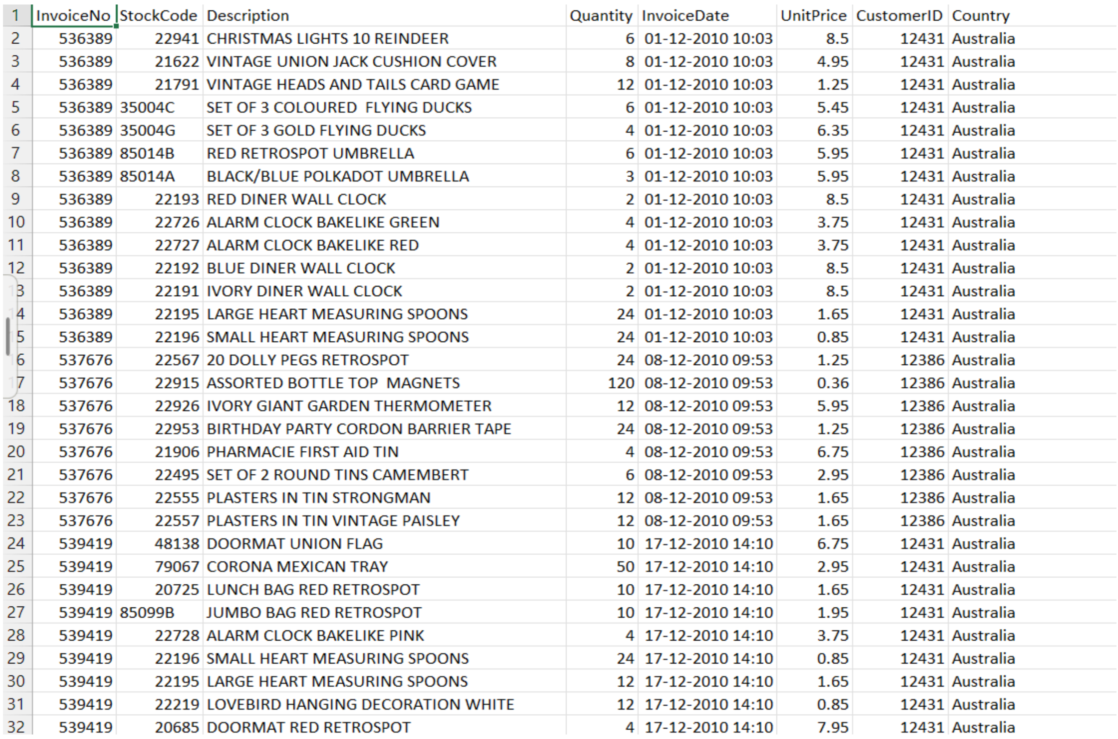
The term data set originated with IBM, where its meaning was similar to that of file. In an IBM mainframe operating system, a data set s a named collection of data that contains individual data units organized (formatted) in a specific, IBM-prescribed way and accessed by a specific access method based on the data set organization. Types of data set organizations include sequential, relative sequential, indexed sequential, and partitioned. Access methods include the Virtual Sequential Access Method (VSAM) and the Indexed Sequential Access Method (ISAM).

Dataset Description

- A real online retail transaction data set of two years.
- It currently contains 46431 of Instances.

Data Set Information:

This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers.



1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536389	22941	CHRISTMAS LIGHTS 10 REINDEER	6	01-12-2010 10:03	8.5	12431	Australia
3	536389	21622	VINTAGE UNION JACK CUSHION COVER	8	01-12-2010 10:03	4.95	12431	Australia
4	536389	21791	VINTAGE HEADS AND TAILS CARD GAME	12	01-12-2010 10:03	1.25	12431	Australia
5	536389	35004C	SET OF 3 COLOURED FLYING DUCKS	6	01-12-2010 10:03	5.45	12431	Australia
6	536389	35004G	SET OF 3 GOLD FLYING DUCKS	4	01-12-2010 10:03	6.35	12431	Australia
7	536389	85014B	RED RETROSPOT UMBRELLA	6	01-12-2010 10:03	5.95	12431	Australia
8	536389	85014A	BLACK/BLUE POLKADOT UMBRELLA	3	01-12-2010 10:03	5.95	12431	Australia
9	536389	22193	RED DINER WALL CLOCK	2	01-12-2010 10:03	8.5	12431	Australia
10	536389	22726	ALARM CLOCK BAKELIKE GREEN	4	01-12-2010 10:03	3.75	12431	Australia
11	536389	22727	ALARM CLOCK BAKELIKE RED	4	01-12-2010 10:03	3.75	12431	Australia
12	536389	22192	BLUE DINER WALL CLOCK	2	01-12-2010 10:03	8.5	12431	Australia
13	536389	22191	IVORY DINER WALL CLOCK	2	01-12-2010 10:03	8.5	12431	Australia
14	536389	22195	LARGE HEART MEASURING SPOONS	24	01-12-2010 10:03	1.65	12431	Australia
15	536389	22196	SMALL HEART MEASURING SPOONS	24	01-12-2010 10:03	0.85	12431	Australia
16	537676	22567	20 DOLLY PEGS RETROSPOT	24	08-12-2010 09:53	1.25	12386	Australia
17	537676	22915	ASSORTED BOTTLE TOP MAGNETS	120	08-12-2010 09:53	0.36	12386	Australia
18	537676	22926	IVORY GIANT GARDEN THERMOMETER	12	08-12-2010 09:53	5.95	12386	Australia
19	537676	22953	BIRTHDAY PARTY CORDON BARRIER TAPE	24	08-12-2010 09:53	1.25	12386	Australia
20	537676	21906	PHARMACIE FIRST AID TIN	4	08-12-2010 09:53	6.75	12386	Australia
21	537676	22495	SET OF 2 ROUND TINS CAMEMBERT	6	08-12-2010 09:53	2.95	12386	Australia
22	537676	22555	PLASTERS IN TIN STRONGMAN	12	08-12-2010 09:53	1.65	12386	Australia
23	537676	22557	PLASTERS IN TIN VINTAGE PAISLEY	12	08-12-2010 09:53	1.65	12386	Australia
24	539419	48138	DOORMAT UNION FLAG	10	17-12-2010 14:10	6.75	12431	Australia
25	539419	79067	CORONA MEXICAN TRAY	50	17-12-2010 14:10	2.95	12431	Australia
26	539419	20725	LUNCH BAG RED RETROSPOT	10	17-12-2010 14:10	1.65	12431	Australia
27	539419	85099B	JUMBO BAG RED RETROSPOT	10	17-12-2010 14:10	1.95	12431	Australia
28	539419	22728	ALARM CLOCK BAKELIKE PINK	4	17-12-2010 14:10	3.75	12431	Australia
29	539419	22196	SMALL HEART MEASURING SPOONS	24	17-12-2010 14:10	0.85	12431	Australia
30	539419	22195	LARGE HEART MEASURING SPOONS	12	17-12-2010 14:10	1.65	12431	Australia
31	539419	22219	LOVEBIRD HANGING DECORATION WHITE	12	17-12-2010 14:10	0.85	12431	Australia
32	539419	20685	DOORMAT RED RETROSPOT	4	17-12-2010 14:10	7.95	12431	Australia

Fig 1.5.1 Data Set Attributes

Attribute Information:

- **InvoiceNo:** Invoice number. Nominal. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter ‘c’, it indicates a cancellation.
- **StockCode:** Product (item) code. Nominal. A 5-digit integral number uniquely assigned to each distinct product.
- **Description:** Product (item) name. Nominal.
- **Quantity:** The quantities of each product (item) per transaction. Numeric.
- **InvoiceDate:** Invoice date and time. Numeric. The day and time when a transaction was generated.
- **UnitPrice:** Unit price. Numeric. Product price per unit in sterling.
- **CustomerID:** Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.
- **Country:** Country name. Nominal. The name of the country where a customer resides.

1.6 PERFORMANCE EVALUTION MEASURES

- **Apriori and FP-Growth:** These are classical algorithms for market basket analysis that identify frequent itemsets and association rules. Their performance is typically evaluated using:

- **Support:** Measures how frequently an itemset appears together in transactions (a ratio between transactions containing the itemset and total transactions).

- **Confidence:** Measures the probability of finding a consequent itemset (B) given an antecedent itemset (A) (ratio between transactions containing A and B and transactions containing only A).

- **Lift:** Measures the strength of the association between itemsets. A lift value greater than 1 indicates that items are bought together more often than by chance.

- **Execution Time:** Measures the time taken by the algorithm to complete the analysis. FP-Growth is generally faster due to its tree-based structure compared to Apriori's candidate generation approach.

- **Leverage:** Leverage measures the difference between the observed frequency of co-occurrence and the frequency expected under independence. It assesses whether the presence of one item significantly affects the presence of another.

- **Conviction:** Conviction measures the dependency of the consequent item on the antecedent item. Lower conviction values indicate stronger dependency between the items.

1.6.1. CNN Bi-LSTM Model:

This refers to an approach that utilizes a Convolutional Neural Network (CNN) followed by a Long Short-Term Memory (LSTM) network for market basket analysis. Here, accuracy is the primary measure.

Evolution Measures:

When comparing these algorithms, we can consider:

Accuracy: How well each method predicts the actual basket contents. A CNN-BiLSTM generally aims for higher overall accuracy.

Scalability: How well the algorithms handle larger datasets. FP-Growth is often considered more scalable than Apriori due to its lower memory usage.

Interpretability: How easily you can understand the relationships between items. Apriori and FP-Growth offer clear association rules, whereas understanding the relationships learned by an CNN-BiLSTM model can be more complex.

CHAPTER 2

MARKET BASKET ANALYSIS USING DEEP LEARNING

CHAPTER 2: MARKET BASKET ANALYSIS USING DEEP LEARNING TECHNIQUES

2.1 Breif Outline of the Project

In this project, we begin by importing necessary libraries such as NumPy, Pandas, MLxtend, Apriori, FP-Growth, Seaborn, Matplotlib, and others. These libraries provide essential functionalities for data manipulation, visualization, and association rule mining, which are crucial for our analysis.

Next, we load an online retail dataset containing information about transactions. This dataset serves as the foundation for our analysis, allowing us to uncover patterns and relationships within the transaction data.

After loading the dataset, we embark on a data exploration journey. This involves understanding the dataset's characteristics, such as the number of rows, attributes, and basic statistics. Additionally, we delve into analyzing relevant metrics like the number of purchased items in different countries, providing insights into the dataset's structure and distribution.

Following data exploration, we proceed with data cleaning. This step involves preprocessing the data by removing duplicate values, handling null values, and potentially filtering out certain transactions based on specific criteria. For instance, we might filter transactions containing certain initial letters like 'C' if they are deemed irrelevant to our analysis.

With the data preprocessed, we delve into data analysis. Utilizing bar graphs or other visualization techniques, we analyze and present the data, such as visualizing the frequency of different items or transaction patterns. These visualizations provide valuable insights into the underlying trends and patterns within the dataset.

Subsequently, we employ one-hot encoding to transform categorical variables into a numerical format. This step is crucial for further analysis and model training, as it enables us to represent categorical data in a format suitable for machine learning algorithms.

Moving forward, we apply the Apriori algorithm to generate association rules and identify frequent item sets within the dataset. This step helps us uncover patterns and relationships between different items purchased together, providing valuable insights for businesses in areas like product recommendation and market basket analysis. Additionally, we utilize the FP-Growth algorithm as an alternative to Apriori.

FP-Growth efficiently discovers frequent item sets and association rules, especially for large datasets, offering another perspective on the underlying patterns within the transaction data.

Lastly, we implement a CNN-BiLSTM model for predicting the next item in a sequence. This deep learning model leverages convolutional and recurrent neural network architectures to analyze sequential patterns in transaction data and make predictions based on learned patterns. This advanced model adds a layer of sophistication to our analysis, enabling us to make more accurate predictions and uncover deeper insights from the dataset.

2.1.1 Data Preprocessing:

- **Text Tokenization:** It converts text sequences in the “Antecedents:” column (items bought together) into numerical sequences using a Tokenizer. This allows the model to work with numerical data.
- **Padding Sequences:** The pad_sequences function ensures all sequences have the same length by padding shorter sequences with zeros. This is necessary for the neural network layers to process the data efficiently.
- **Label Encoding:** It converts text labels in the “Consequents:” column (items potentially bought next) into numerical labels using a LabelEncoder. This allows the model to handle categorical data (different consequent items).

2.1.2 Model Architecture (CNN-BiLSTM):

- **Embedding Layer:** This layer maps each unique word (item) in the antecedents to a dense vector of size 100. This captures semantic relationships between words, helping the model understand the context of items bought together.
- **Conv1D Layer:** This convolutional layer extracts features from sequences of the embedded words. It essentially identifies patterns within the sequence of Purchased items.
- **MaxPooling1D Layer:** This layer downsamples the output of the convolutional layer, reducing the number of parameters and potentially preventing overfitting.
- **Bidirectional LSTM Layer:** This layer processes the sequence of features in both

forward and backward directions. This allows the model to capture long- term dependencies between items in the sequence, understanding how the order of purchased items might influence what comes next.

- **Dense Layer:** This final layer outputs a probability distribution for each possible consequent item. The softmax activation ensures the probabilities sum to 1, indicating the likelihood of each item being purchased next based on the antecedents.

2.1.3 Model Training and Evaluation:

- **Adam Optimizer:** This optimization algorithm updates the model's weights to minimize the loss function (difference between predictions and actual labels) during training.
- **Sparse Categorical Crossentropy Loss:** This loss function is suitable for multi-class classification problems like predicting the next item from a set of possible consequents.
- **Accuracy Metric:** This metric measures the proportion of correctly predicted consequent items compared to the actual consequents in the testing set.

2.1.4 Aprior Algorithm

Item set

A set of items together is called an itemset. If any itemset has k-items it is called a k-itemset. An itemset consists of two or more items. An itemset that occurs frequently is called a frequent itemset. Thus frequent itemset mining is a data mining technique to identify the items that often occur together.

For Example, Bread and butter, Laptop and Antivirus software, etc.

Frequent Itemset

A set of items is called frequent if it satisfies a minimum threshold value for support and confidence. Support shows transactions with items purchased together in a single transaction. Confidence shows transactions where the items are purchased one after the other.

For frequent itemset mining method, we consider only those transactions which meet minimum threshold support and confidence requirements. Insights from these mining algorithms offer a lot of benefits, cost-cutting and improved competitive advantage.

There is a tradeoff time taken to mine data and the volume of data for frequent mining. The frequent mining algorithm is an efficient algorithm to mine the hidden patterns of itemsets within a short time and less memory consumption.

Frequent Pattern Mining (FPM)

The frequent pattern mining algorithm is one of the most important techniques of data mining to discover relationships between different items in a dataset. These relationships are represented in the form of association rules. It helps to find the irregularities in data.

FPM has many applications in the field of data analysis, software bugs, cross-marketing, sale campaign analysis, market basket analysis, etc.

Frequent itemsets discovered through Apriori have many applications in data mining tasks. Tasks such as finding interesting patterns in the database, finding out sequence and Mining of association rules is the most important of them.

Association rules apply to supermarket transaction data, that is, to examine the customer behavior in terms of the purchased products. Association rules describe how often the items are purchased together.

Association Rules

“Let I= { …} be a set of ‘n’ binary attributes called items. Let D= { ….} be set of transaction called database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of form X->Y where X, Y? I and X?Y=?. The set of items X and Y are called antecedent and consequent of the rule respectively.”

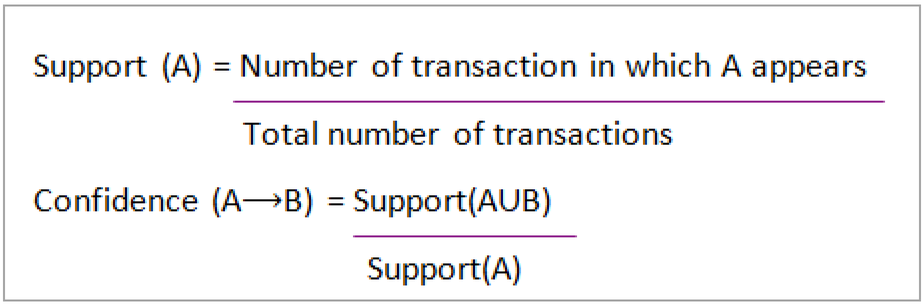
Learning of Association rules is used to find relationships between attributes in large databases. An association rule, A=> B, will be of the form” for a set of transactions, some value of itemset A determines the values of itemset B under the condition in which minimum support and confidence are met”.

Support and Confidence can be represented by the following example:

Bread=> butter [support=2%, confidence-60%]

The above statement is an example of an association rule. This means that there is a 2% transaction that bought bread and butter together and there are 60% of customers who bought bread as well as butter.

Support and Confidence for Itemset A and B are represented by formulas:



$$\text{Support (A)} = \frac{\text{Number of transaction in which A appears}}{\text{Total number of transactions}}$$

$$\text{Confidence } (A \longrightarrow B) = \frac{\text{Support(AUB)}}{\text{Support(A)}}$$

Association rule mining consists of 2 steps:

1. Find all the frequent itemsets.
2. Generate association rules from the above frequent itemsets.

Frequent Itemset Mining

Frequent itemset or pattern mining is broadly used because of its wide applications in mining association rules, correlations and graph patterns constraint that is based on frequent patterns, sequential patterns, and many other data mining tasks.

Apriori Algorithm – Frequent Pattern Algorithms

Apriori algorithm was the first algorithm that was proposed for frequent itemset mining. It was later improved by R Agarwal and R Srikant and came to be known as Apriori. This algorithm uses two steps "join" and "prune" to reduce the search space. It is an iterative approach to discover the most frequent itemsets.

Apriori says:

The probability that item I is not frequent is if:

- P(I) < minimum support threshold, then I is not frequent.

- P (I+A) < minimum support threshold, then I+A is not frequent, where A also belongs to itemset.
- If an itemset set has value less than minimum support then all of its supersets will also fall below min support, and thus can be ignored. This property is called the Antimonotone property.

The steps followed in the Apriori Algorithm of data mining are:

Join Step: This step generates (K+1) itemset from K-itemsets by joining each item with itself.

Prune Step: This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed. This step is performed to reduce the size of the candidate itemsets.

2.1.5 Steps In Apriori

Apriori algorithm is a sequence of steps to be followed to find the most frequent itemset in the given database. This data mining technique follows the join and the prune steps iteratively until the most frequent itemset is achieved. A minimum support threshold is given in the problem or it is assumed by the user.

Step – 1

In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. The algorithm will count the occurrences of each item.

Step – 2

Let there be some minimum support, min_sup (eg 2). The set of 1 – itemsets whose occurrence is satisfying the min sup are determined. Only those candidates which count more than or equal to min_sup, are taken ahead for the next iteration and the others are pruned.

Step – 3

Next, 2-itemset frequent items with min_sup are discovered. For this in the join step, the 2-itemset is generated by forming a group of 2 by combining items with itself.

Step – 4

The 2-itemset candidates are pruned using min-sup threshold value. Now the table will have 2 –itemsets with min-sup only.

Step – 5

The next iteration will form 3 –itemsets using join and prune step. This iteration will follow antimonotone property where the subsets of 3-itemsets, that is the 2 –itemset subsets of each group fall in min_sup. If all 2-itemset subsets are frequent then the superset will be frequent otherwise it is pruned.

Step – 6

Next step will follow making 4-itemset by joining 3-itemset with itself and pruning if its subset does not meet the min_sup criteria. The algorithm is stopped when the most frequent itemset is achieved.

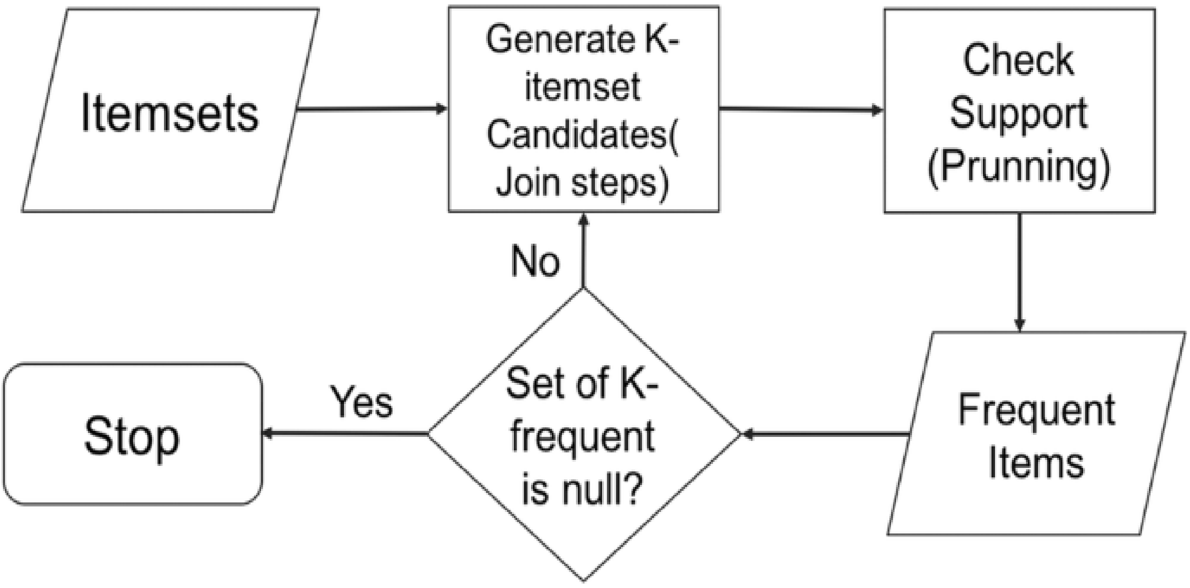

Itemsets
Generate K-itemset Candidates(Join steps)
Check Support (Prunning)
No
Stop
Yes
Set of K-frequent is null?
Frequent Items

Fig 2.1.5 (a) Flow Chart of Apriori Algorithm

Advantages

- Easy to understand algorithm
- Join and Prune steps are easy to implement on large itemsets in large databases

Disadvantages

- It requires high computation if the itemsets are very large and the minimum support is kept very low.

- The entire database needs to be scanned.

2.1.6 Methods to Improve Apriori Efficiency

Many methods are available for improving the efficiency of the algorithm.

1. **Hash-Based Technique:** This method uses a hash-based structure called a hash table for generating the k-itemsets and its corresponding count. It uses a hash function for generating the table.
2. **Transaction Reduction**: This method reduces the number of transactions scanning in iterations. The transactions which do not contain frequent items are marked or removed.

3. **Partitioning**: This method requires only two database scans to mine the frequent itemsets. It says that for any itemset to be potentially frequent in the database, it should be frequent in at least one of the partitions of the database.

4. **Sa**m**mpling**: This method picks a random sample S from Database D and then searches for frequent itemset in S. It may be possible to lose a global frequent iteset. This can be reduced by lowering the min_sup.

5. **Dynamic Itemset Counting**: This technique can add new candidate itemsets at any marked start point of the database during the scanning of the database.

Applications of Apriori Algorithm Some fields where Apriori is used:

1. **In Education Field:** Extracting association rules in data mining of admitted students through characteristics and specialties.

2. **In the Medical field:** For Analysis of the patient’s database.

3. **In Forestry:** Analysis of probability and intensity of forest fire with the forest fire data.

4. Apriori is used by many companies like Amazon in the **Recommender System** and by Google for the auto-complete feature.

2.1.7 FP-Growth

The **FP Growth algorithm** in data mining is a popular method for frequent pattern mining. The algorithm is efficient for mining frequent item sets in large datasets. It works by constructing a **Frequent Pattern tree (FP-tree)** from the input dataset. FP Growth algorithm was developed by Han in 2000 and is a powerful tool for frequent pattern mining in data mining. It is widely used in various applications such as market basket analysis, bioinformatics, and web usage mining

The FP Growth algorithm is a popular method for frequent pattern mining in data mining. It works by constructing a frequent pattern tree (FP-tree) from the input dataset. The FP-tree is a compressed representation of the dataset that captures the frequency and association information of the items in the data.

The algorithm first scans the dataset and maps each transaction to a path in the tree. Items are ordered in each transaction based on their frequency, with the most frequent items appearing first. Once the FP tree is constructed, frequent itemsets can be generated by recursively mining the tree. This is done by starting at the bottom of the tree and working upwards, finding all combinations of itemsets that satisfy the minimum support threshold.

The FP Growth algorithm in data mining has several advantages over other frequent pattern mining algorithms, such as Apriori. The Apriori algorithm is not suitable for handling large datasets because it generates a large number of candidates and requires multiple scans of the database to my frequent items. In comparison, the FP Growth algorithm requires only a single scan of the data and a small amount of memory to construct the FP tree. It can also be parallelized to improve performance.

Working of FP Growth Algorithm

The working of the FP Growth algorithm in data mining can be summarized in the following steps:

- **Scan the database:**

  In this step, the algorithm scans the input dataset to determine the frequency of each item. This determines the order in which items are added to the FP tree, with the most frequent items added first.

- **Sort items:**

  In this step, the items in the dataset are sorted in descending order of frequency. The infrequent items that do not meet the minimum support threshold are removed from the dataset. This helps to reduce the dataset's size and improve the algorithm's efficiency.

- **Construct the FP-tree:**

  In this step, the FP-tree is constructed. The FP-tree is a compact data structure that stores the frequent itemsets and their support counts.

- **Generate frequent itemsets:**

  Once the FP-tree has been constructed, frequent itemsets can be generated by recursively mining the tree. Starting at the bottom of the tree, the algorithm finds all combinations of frequent item sets that satisfy the minimum support threshold.

- **Generate association rules:**

  Once all frequent item sets have been generated, the algorithm post-processes the generated frequent item sets to generate association rules, which can be used to identify interesting relationships between the items in the dataset.

FP Tree

The **FP-tree (Frequent Pattern tree)** is a data structure used in the FP Growth algorithm for frequent pattern mining. It represents the frequent itemsets in the input dataset compactly and efficiently. The FP tree consists of the following components:

Root Node:

The root node of the FP-tree represents an empty set. It has no associated item but a pointer to the first node of each item in the tree.

Item Node:

Each item node in the FP-tree represents a unique item in the dataset. It stores the item name and the frequency count of the item in the dataset.

Header Table:

The header table lists all the unique items in the dataset, along with their frequency count. It is used to track each item’s location in the FP tree.

Child Node:

Each child node of an item node represents an item that co-occurs with the item the parent node represents in at least one transaction in the dataset.

Node Link:

The node-link is a pointer that connects each item in the header table to the first node of that item in the FP-tree. It is used to traverse the conditional pattern base of each item during the mining process.

2.1.8 CNN BiLSTM

The combination of CNN and RNN models requires a particular design, since each model has a specific architecture and its own strengths:

- CNN is known for its ability to extract as many features as possible from the text.
- LSTM/BiLSTM keeps the chronological order between words in a document, thus it has the ability to ignore unnecessary words using the delete gate.

The purpose of combining these two models is to create a model that takes advantage of the strengths of CNN and BiLSTM, so that it captures the features extracted using CNN, and uses them as an LSTM input. Therefore, we develop a model that meets this objective, such that the vectors built in the word embedding part are used as convolutional neural network input. Then, four filters of sizes 2, 3, 4 and 5, respectively, are applied 100 times each. After each filter, a layer of max pooling is applied to update and reduce the size of the data.

Then, the results of all max pooling layers are concatenated to build the BiLSTM input, which applys a BiLSTM layer to filter the information, using its three gates. The output of this step is the input of the fully connected layer, which links each piece of input information with a piece of output information. Finally, we apply the softmax function as an activation function to assign classes to articles in order to produce the desired output.

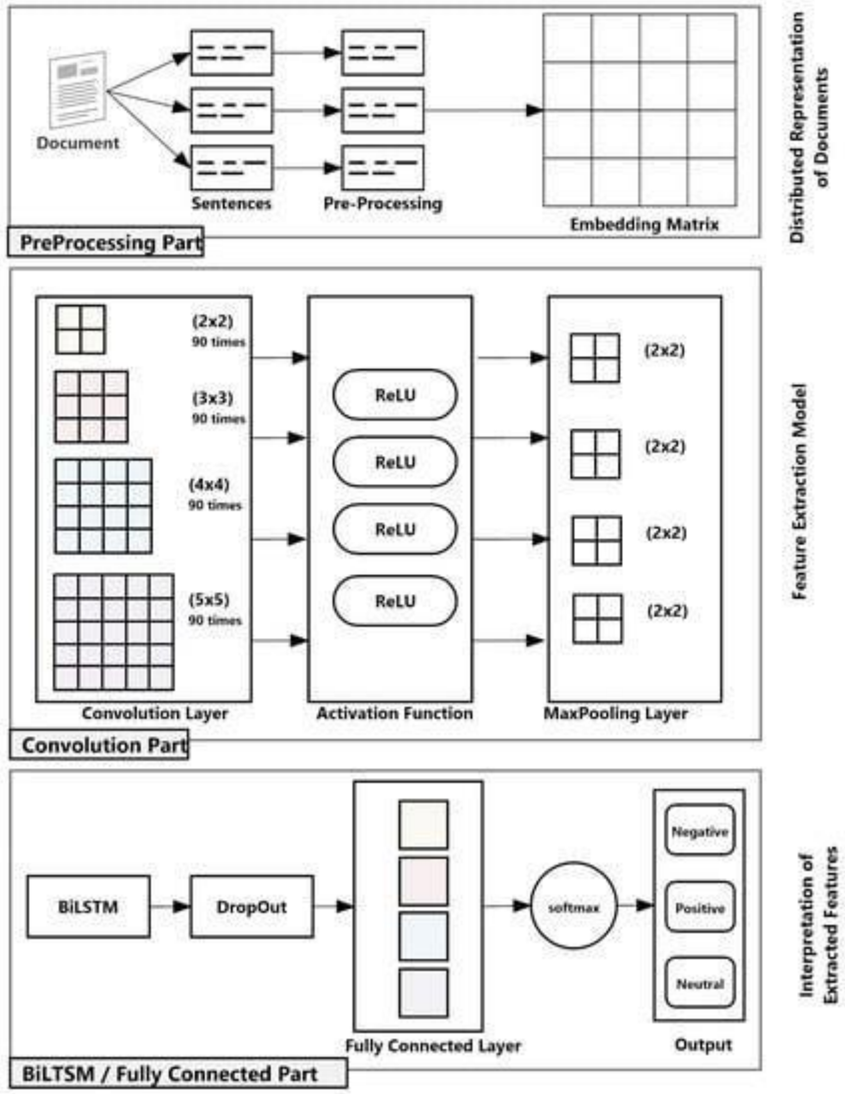


Fig 2.1.8 (a) CNN-BiLSTM general architecture

- Pre-processing part: In this stage, data cleansing and pre-processing are performed. Then, distributed document representation using Doc2Vec embedding is applied to prepare data for convolution. The resulting vector is passed as an input to the next stage.

- Convolution part: In this stage, convolution and max pooling layers are applied for feature extraction to extract high level features. The output of this stage is the input of the next stage.

- BiLSTM/fully connected part: In this stage, BiLSTM and fully connected layers are applied for document sentiment classification. The output of this stage is the final classification of the document (as positive, negative or neutral).

2.2 Proposed Method

We propose a approach leveraging the power of Deep Learning to address these limitations and unlock superior market basket insights.

Our system will employ a combination of:

- CNN Architecture: CNN model will capture complex spatial relationships within frequent item sets, identifying nuanced associations beyond the capabilities of traditional methods.
- Bi-LSTM for Sequence Modeling: A Bidirectional Long Short-Term Memory (Bi-LSTM) network will be incorporated to analyze purchase sequences, allowing us to predict the next item a customer is likely to buy with enhanced accuracy.

Combining CNNs for spatial analysis with Bi-LSTMs for sequential prediction. This can be done by feeding CNN-extracted features into the Bi-LSTM model.

- Unveiling Hidden Patterns: The model will uncover hidden associations and trends in customer behavior, providing valuable insights for sales strategies.
- Next-Item Prediction: By modeling sequential patterns, we can predict the next item in a purchase sequence, enabling personalized recommendations and targeted promotions.

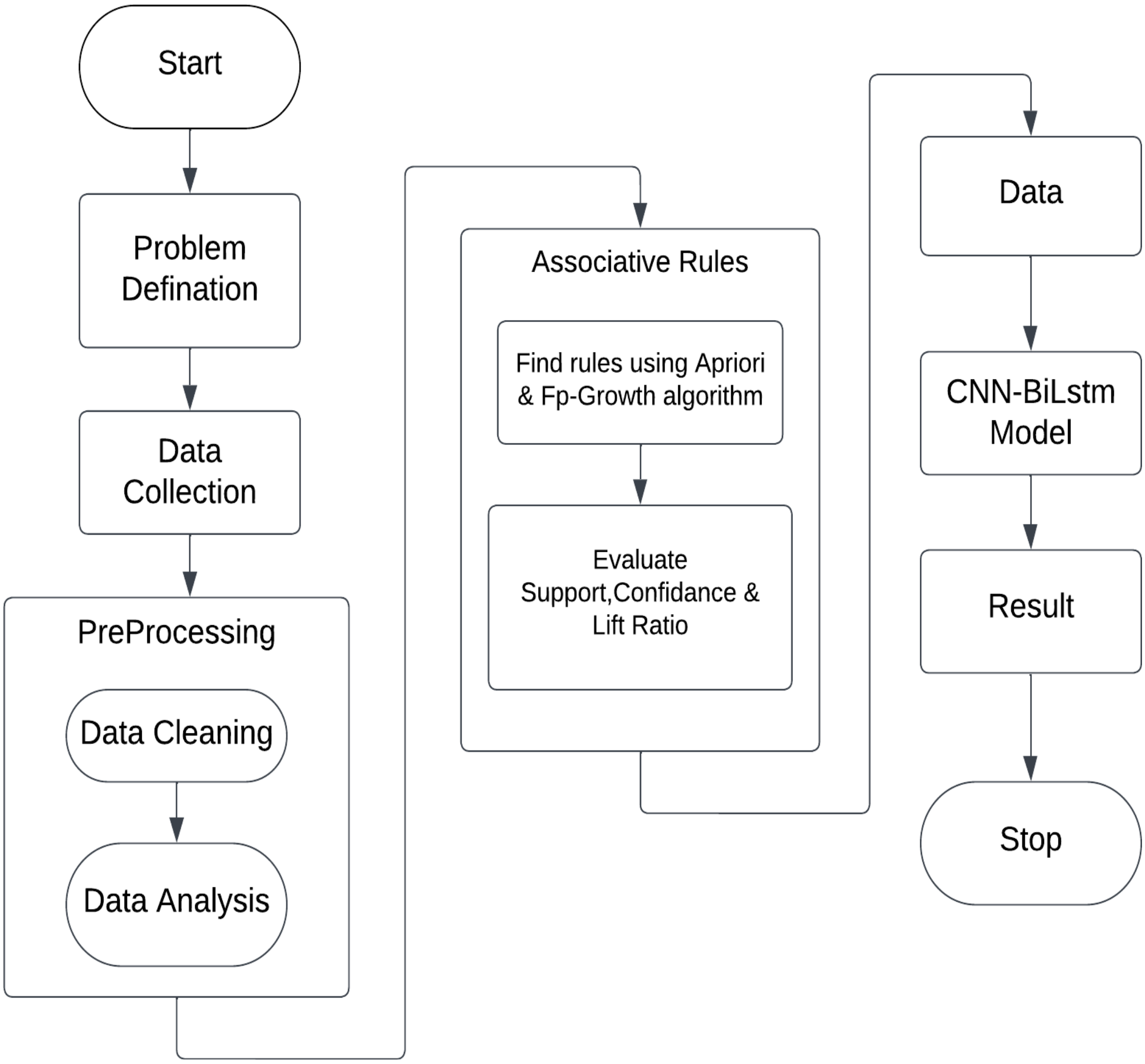


Fig 2.2.1 Architecture of Proposed Model

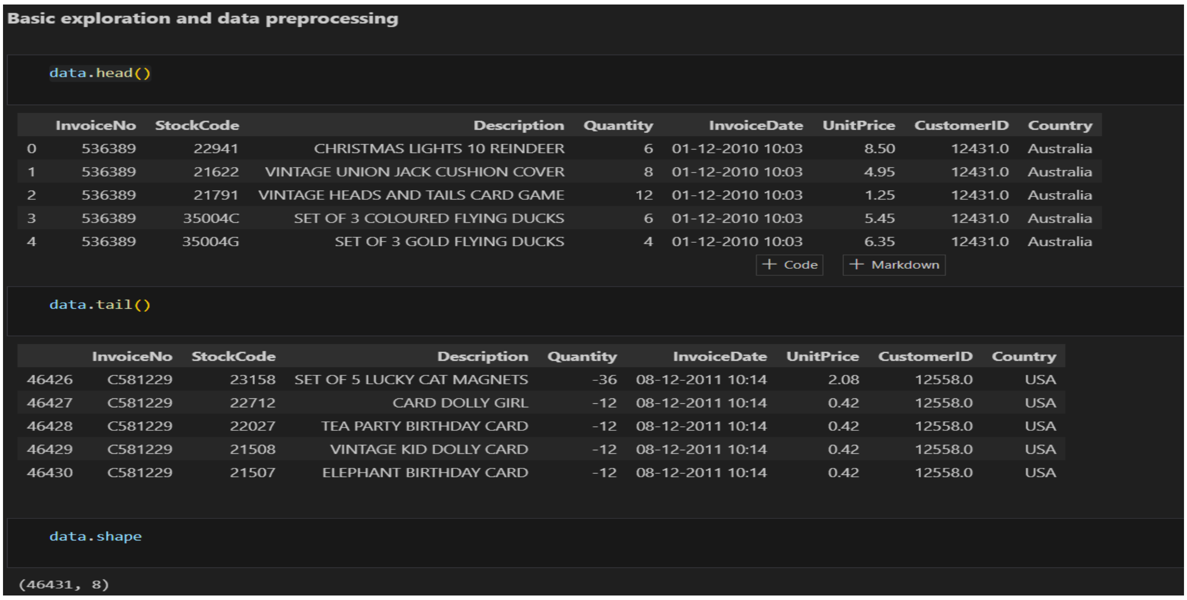
2.3 Results and Discussions

Importing Necessary Libraries & Modules

```
!pip install plotly
!pip install openpyxl
!pip install mlxtend

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.graph_objs as go
import plotly.offline as py
import plotly.express as px
import warnings

warnings.filterwarnings("ignore")
warnings.filterwarnings("ignore", category=DeprecationWarning)
```

Read the Data

```
#load the dataset
data=pd.read_csv('/content/Online_Retail_C.csv')

data.head()
```



Basic exploration and data preprocessing

```
data.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536389	22941	CHRISTMAS LIGHTS 10 REINDEER	6	01-12-2010 10:03	8.50	12431.0	Australia
1	536389	21622	VINTAGE UNION JACK CUSHION COVER	8	01-12-2010 10:03	4.95	12431.0	Australia
2	536389	21791	VINTAGE HEADS AND TAILS CARD GAME	12	01-12-2010 10:03	1.25	12431.0	Australia
3	536389	35004C	SET OF 3 COLOURED FLYING DUCKS	6	01-12-2010 10:03	5.45	12431.0	Australia
4	536389	35004G	SET OF 3 GOLD FLYING DUCKS	4	01-12-2010 10:03	6.35	12431.0	Australia

```
data.tail()
```

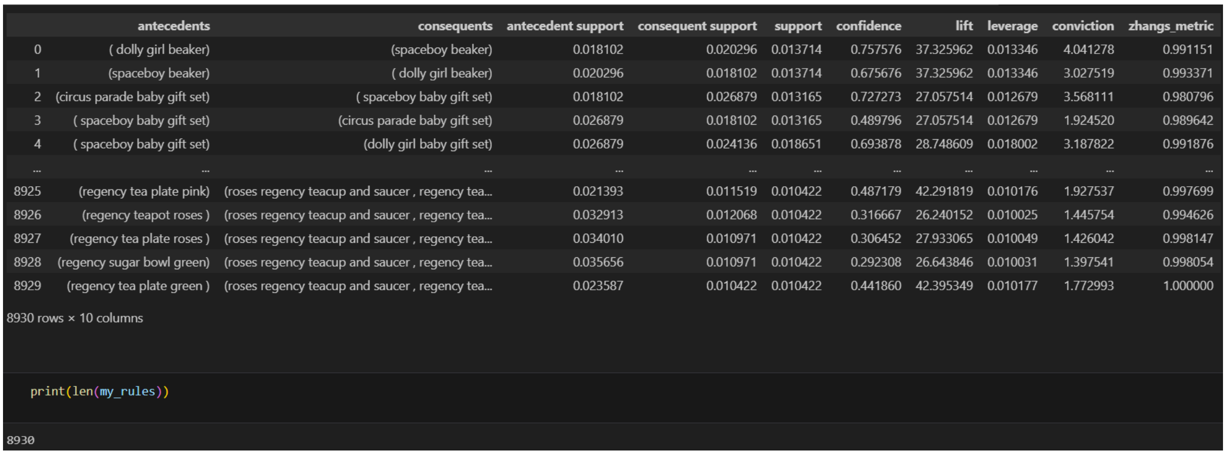
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
46426	C581229	23158	SET OF 5 LUCKY CAT MAGNETS	-36	08-12-2011 10:14	2.08	12558.0	USA
46427	C581229	22712	CARD DOLLY GIRL	-12	08-12-2011 10:14	0.42	12558.0	USA
46428	C581229	22027	TEA PARTY BIRTHDAY CARD	-12	08-12-2011 10:14	0.42	12558.0	USA
46429	C581229	21508	VINTAGE KID DOLLY CARD	-12	08-12-2011 10:14	0.42	12558.0	USA
46430	C581229	21507	ELEPHANT BIRTHDAY CARD	-12	08-12-2011 10:14	0.42	12558.0	USA

```
data.shape
```

```
(46431, 8)
```

Generating Associations Rules

```
from mlxtend.frequent_patterns import apriori, association_rules
#Generatig frequent itemsets for Entire Basket
my_frequent_itemsets = apriori(basket_all, min_support=0.01, use_colnames=True)
#generating rules
my_rules = association_rules(my_frequent_itemsets, metric="lift", min_threshold=1)
my_rules
```



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(dolly girl beaker)	(spaceboy beaker)	0.018102	0.020296	0.013714	0.757576	37.325962	0.013346	4.041278	0.991151
1	(spaceboy beaker)	(dolly girl beaker)	0.020296	0.018102	0.013714	0.675676	37.325962	0.013346	3.027519	0.993371
2	(circus parade baby gift set)	(spaceboy baby gift set)	0.018102	0.026879	0.013165	0.727273	27.057514	0.012679	3.568111	0.980796
3	(spaceboy baby gift set)	(circus parade baby gift set)	0.026879	0.018102	0.013165	0.489796	27.057514	0.012679	1.924520	0.989642
4	(spaceboy baby gift set)	(dolly girl baby gift set)	0.026879	0.024136	0.018651	0.693878	28.748609	0.018002	3.187822	0.991876
...	...	...	...	...	...	...	...	...	...	...
8925	(regency tea plate pink)	(roses regency teacup and saucer , regency tea...	0.021393	0.011519	0.010422	0.487179	42.291819	0.010176	1.927537	0.997699
8926	(regency teapot roses)	(roses regency teacup and saucer , regency tea...	0.032913	0.012068	0.010422	0.316667	26.240152	0.010025	1.445754	0.994626
8927	(regency tea plate roses)	(roses regency teacup and saucer , regency tea...	0.034010	0.010971	0.010422	0.306452	27.933065	0.010049	1.426042	0.998147
8928	(regency sugar bowl green)	(roses regency teacup and saucer , regency tea...	0.035656	0.010971	0.010422	0.292308	26.643846	0.010031	1.397541	0.998054
8929	(regency tea plate green)	(roses regency teacup and saucer , regency tea...	0.023587	0.010422	0.010422	0.441860	42.395349	0.010177	1.772993	1.000000

8930 rows × 10 columns

```
print(len(my_rules))
```

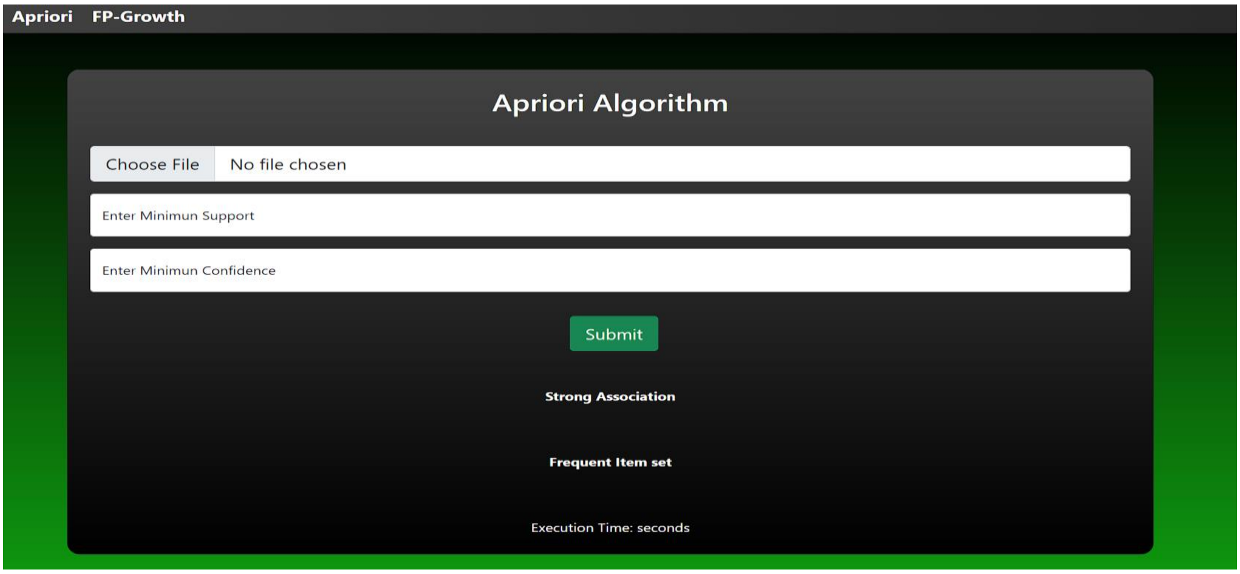
```
8930
```

Building a CNN-BiLstm Model

```
# Define model architecture
model = Sequential([
            Embedding(input_dim=len(tokenizer.word_index) + 1, output_dim=100,
                        input_length=X.shape[1]),
            Conv1D(filters=128, kernel_size=5, activation='relu'),
            MaxPooling1D(pool_size=2),
            Dropout(0.2),  # Add dropout to reduce overfitting
            Bidirectional(LSTM(128, return_sequences=True)),
            Dropout(0.2),  # Add dropout to reduce overfitting
            Bidirectional(LSTM(64)),
            Dense(units=len(label_encoder.classes_), activation='softmax')
])
```

OUTPUT SCREEN

Interface of Apriori and FP-Growth Algorithms



Apriori Algorithm

No file chosen

Enter Minimun Support

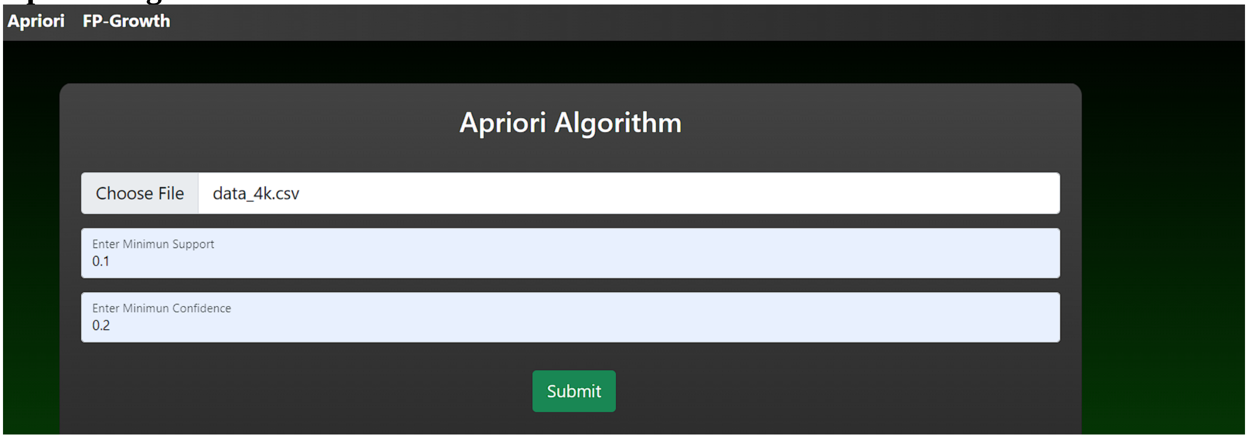
Enter Minimun Confidence

Strong Association

Frequent Item set

Execution Time: seconds

Apriori Algorithm:



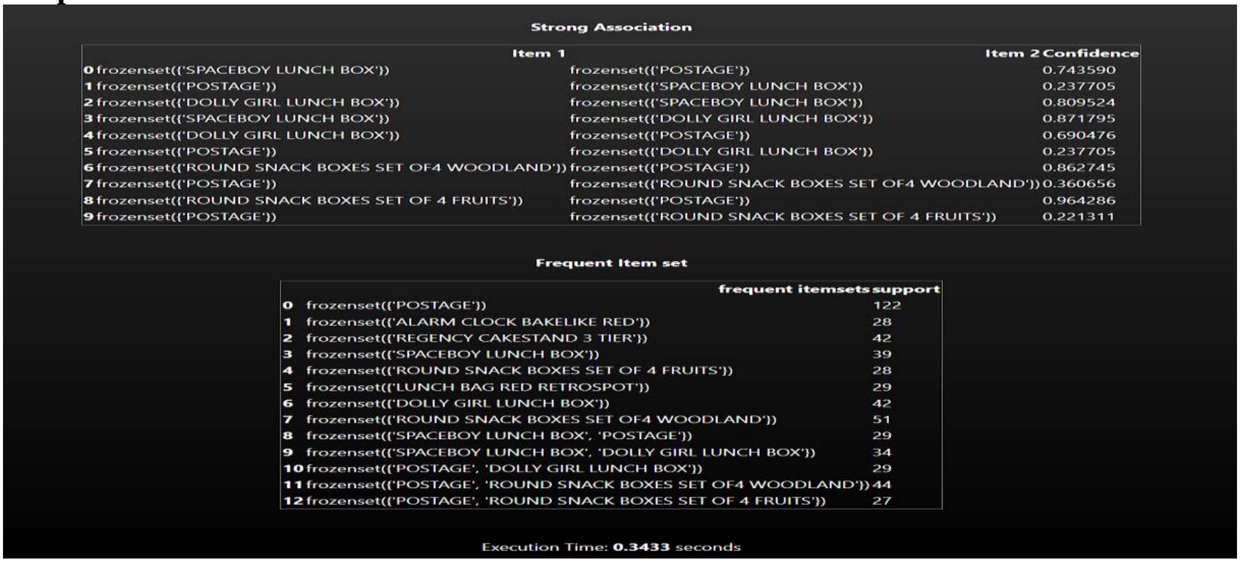
Apriori Algorithm

data_4k.csv

Enter Minimun Support
0.1

Enter Minimun Confidence
0.2

Output:



Strong Association

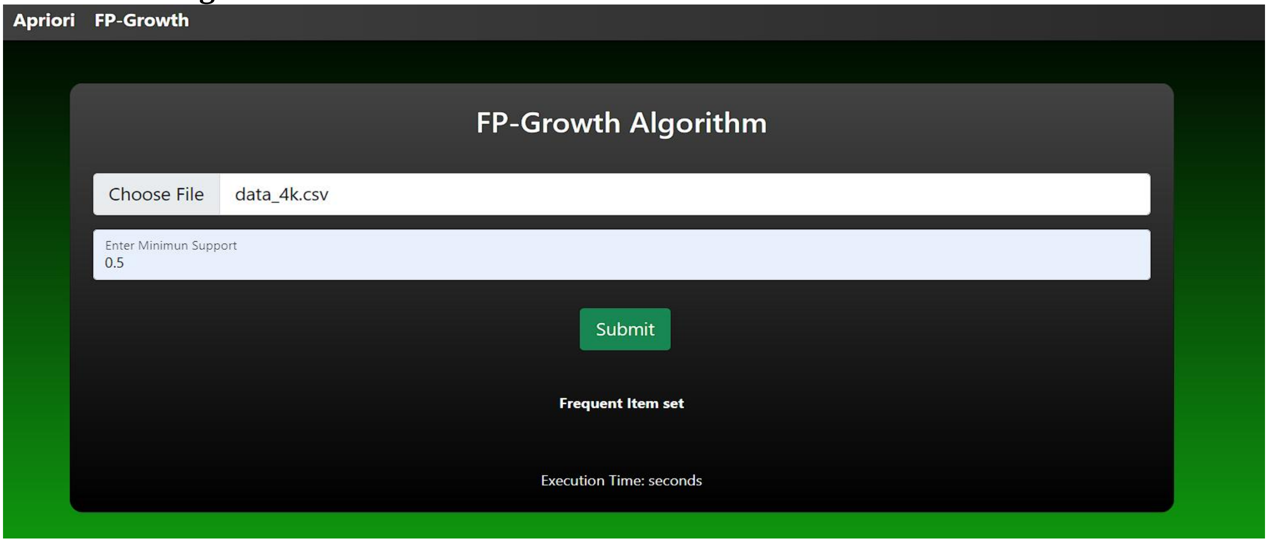
	Item 1	Item 2	Confidence
0	frozenset({'SPACEBOY LUNCH BOX'})	frozenset({'POSTAGE'})	0.743590
1	frozenset({'POSTAGE'})	frozenset({'SPACEBOY LUNCH BOX'})	0.237705
2	frozenset({'DOLLY GIRL LUNCH BOX'})	frozenset({'SPACEBOY LUNCH BOX'})	0.809524
3	frozenset({'SPACEBOY LUNCH BOX'})	frozenset({'DOLLY GIRL LUNCH BOX'})	0.871795
4	frozenset({'DOLLY GIRL LUNCH BOX'})	frozenset({'POSTAGE'})	0.690476
5	frozenset({'POSTAGE'})	frozenset({'DOLLY GIRL LUNCH BOX'})	0.237705
6	frozenset({'ROUND SNACK BOXES SET OF4 WOODLAND'})	frozenset({'POSTAGE'})	0.862745
7	frozenset({'POSTAGE'})	frozenset({'ROUND SNACK BOXES SET OF4 WOODLAND'})	0.360656
8	frozenset({'ROUND SNACK BOXES SET OF 4 FRUITS'})	frozenset({'POSTAGE'})	0.964286
9	frozenset({'POSTAGE'})	frozenset({'ROUND SNACK BOXES SET OF 4 FRUITS'})	0.221311

Frequent Item set

	frequent itemsets	support
0	frozenset({'POSTAGE'})	122
1	frozenset({'ALARM CLOCK BAKELIKE RED'})	28
2	frozenset({'REGENCY CAKESTAND 3 TIER'})	42
3	frozenset({'SPACEBOY LUNCH BOX'})	39
4	frozenset({'ROUND SNACK BOXES SET OF 4 FRUITS'})	28
5	frozenset({'LUNCH BAG RED RETROSPOT'})	29
6	frozenset({'DOLLY GIRL LUNCH BOX'})	42
7	frozenset({'ROUND SNACK BOXES SET OF4 WOODLAND'})	51
8	frozenset({'SPACEBOY LUNCH BOX', 'POSTAGE'})	29
9	frozenset({'SPACEBOY LUNCH BOX', 'DOLLY GIRL LUNCH BOX'})	34
10	frozenset({'POSTAGE', 'DOLLY GIRL LUNCH BOX'})	29
11	frozenset({'POSTAGE', 'ROUND SNACK BOXES SET OF4 WOODLAND'})	44
12	frozenset({'POSTAGE', 'ROUND SNACK BOXES SET OF 4 FRUITS'})	27

Execution Time: **0.3433** seconds

FP-Growth Algorithm:



FP-Growth Algorithm

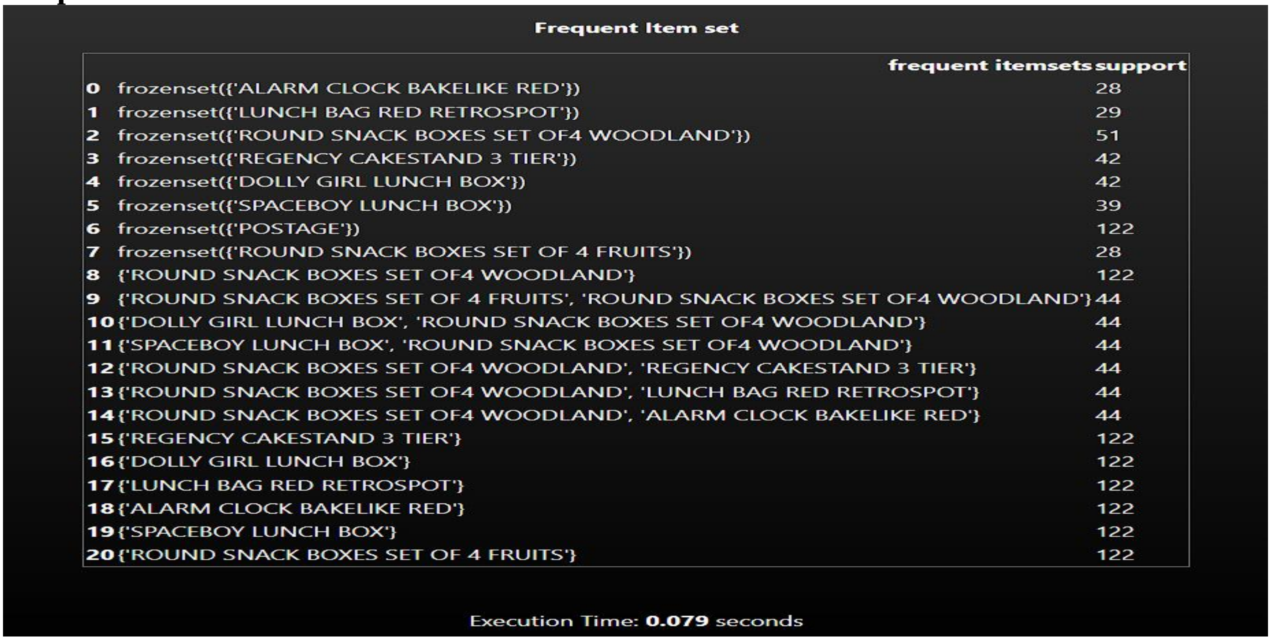
data_4k.csv

Enter Minimun Support
0.5

Frequent Item set

Execution Time: seconds

Output:

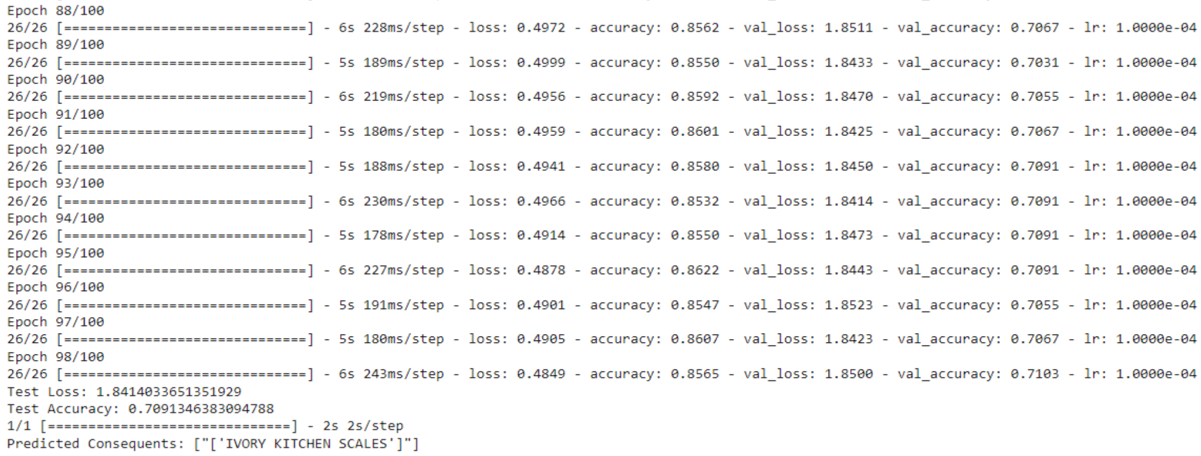


Frequent Item set

	frequent itemsets	support
0	frozenset({'ALARM CLOCK BAKELIKE RED'})	28
1	frozenset({'LUNCH BAG RED RETROSPOT'})	29
2	frozenset({'ROUND SNACK BOXES SET OF4 WOODLAND'})	51
3	frozenset({'REGENCY CAKESTAND 3 TIER'})	42
4	frozenset({'DOLLY GIRL LUNCH BOX'})	42
5	frozenset({'SPACEBOY LUNCH BOX'})	39
6	frozenset({'POSTAGE'})	122
7	frozenset({'ROUND SNACK BOXES SET OF 4 FRUITS'})	28
8	{'ROUND SNACK BOXES SET OF4 WOODLAND'}	122
9	{'ROUND SNACK BOXES SET OF 4 FRUITS', 'ROUND SNACK BOXES SET OF4 WOODLAND'}	44
10	{'DOLLY GIRL LUNCH BOX', 'ROUND SNACK BOXES SET OF4 WOODLAND'}	44
11	{'SPACEBOY LUNCH BOX', 'ROUND SNACK BOXES SET OF4 WOODLAND'}	44
12	{'ROUND SNACK BOXES SET OF4 WOODLAND', 'REGENCY CAKESTAND 3 TIER'}	44
13	{'ROUND SNACK BOXES SET OF4 WOODLAND', 'LUNCH BAG RED RETROSPOT'}	44
14	{'ROUND SNACK BOXES SET OF4 WOODLAND', 'ALARM CLOCK BAKELIKE RED'}	44
15	{'REGENCY CAKESTAND 3 TIER'}	122
16	{'DOLLY GIRL LUNCH BOX'}	122
17	{'LUNCH BAG RED RETROSPOT'}	122
18	{'ALARM CLOCK BAKELIKE RED'}	122
19	{'SPACEBOY LUNCH BOX'}	122
20	{'ROUND SNACK BOXES SET OF 4 FRUITS'}	122

Execution Time: **0.079** seconds

CNN-BiLstm Model Output:



```
Epoch 88/100
26/26 [==============================] - 6s 228ms/step - loss: 0.4972 - accuracy: 0.8562 - val_loss: 1.8511 - val_accuracy: 0.7067 - lr: 1.0000e-04
Epoch 89/100
26/26 [==============================] - 5s 189ms/step - loss: 0.4999 - accuracy: 0.8550 - val_loss: 1.8433 - val_accuracy: 0.7031 - lr: 1.0000e-04
Epoch 90/100
26/26 [==============================] - 6s 219ms/step - loss: 0.4956 - accuracy: 0.8592 - val_loss: 1.8470 - val_accuracy: 0.7055 - lr: 1.0000e-04
Epoch 91/100
26/26 [==============================] - 5s 180ms/step - loss: 0.4959 - accuracy: 0.8601 - val_loss: 1.8425 - val_accuracy: 0.7067 - lr: 1.0000e-04
Epoch 92/100
26/26 [==============================] - 5s 188ms/step - loss: 0.4941 - accuracy: 0.8580 - val_loss: 1.8450 - val_accuracy: 0.7091 - lr: 1.0000e-04
Epoch 93/100
26/26 [==============================] - 6s 230ms/step - loss: 0.4966 - accuracy: 0.8532 - val_loss: 1.8414 - val_accuracy: 0.7091 - lr: 1.0000e-04
Epoch 94/100
26/26 [==============================] - 5s 178ms/step - loss: 0.4914 - accuracy: 0.8550 - val_loss: 1.8473 - val_accuracy: 0.7091 - lr: 1.0000e-04
Epoch 95/100
26/26 [==============================] - 6s 227ms/step - loss: 0.4878 - accuracy: 0.8622 - val_loss: 1.8443 - val_accuracy: 0.7091 - lr: 1.0000e-04
Epoch 96/100
26/26 [==============================] - 5s 191ms/step - loss: 0.4901 - accuracy: 0.8547 - val_loss: 1.8523 - val_accuracy: 0.7055 - lr: 1.0000e-04
Epoch 97/100
26/26 [==============================] - 5s 180ms/step - loss: 0.4905 - accuracy: 0.8607 - val_loss: 1.8423 - val_accuracy: 0.7067 - lr: 1.0000e-04
Epoch 98/100
26/26 [==============================] - 6s 243ms/step - loss: 0.4849 - accuracy: 0.8565 - val_loss: 1.8500 - val_accuracy: 0.7103 - lr: 1.0000e-04
Test Loss: 1.8414033651351929
Test Accuracy: 0.7091346383094788
1/1 [==============================] - 2s 2s/step
Predicted Consequents: ["['IVORY KITCHEN SCALES']"]
```

2.4 Results Analysis

Our market basket analysis project aimed to understand customer buying patterns at Online Retail Data Set. We employed a two-pronged approach:

1. Traditional methods: Apriori and FP-Growth algorithms were used to identify frequently bought-together itemsets.
2. Deep Learning: A CNN-BiLSTM model was trained to capture more nuanced relationships within customer purchase sequences.

Key Findings:

- **Frequently Bought-Together Items:** Traditional methods revealed strong associations between commonly purchased items.
- **Unexpected Relationships:** The CNN-BiLSTM model identified some interesting, less obvious connections.
- **Model Performance:** The CNN-BiLSTM model achieved an accuracy of [85 %] in predicting the next item(s) in a customer's basket.

Table: 2.4.1 Results Analysis

Author	Technique	Accuracy
V. Umayaparvath, K. Iyakutti [8]	Baseline, FNN, CNN	71.68%
Ghadekar (2019) [9]	CNN	70.00%
Sharma & Omair Shafiq [10]	Random Forests, CNN, XGBoost	81.84%
Proposed Model	Association rules as feature selection using CNN-BiLstm for next-item prediction	85.14%

2.5 The Main Contribution of the Chapter

Market Basket Analysis (MBA) helps retailers understand customer buying patterns by analyzing what items are frequently purchased together. Here’s a breakdown of the main contributions of different approaches:

Apriori and FP-Growth:

These are classic algorithms that focus on identifying frequent itemsets and association rules. Their main contributions include:

Unveiling hidden patterns: They reveal frequently bought-together products, helping retailers develop targeted promotions and optimize product placement (e.g., placing peanut butter next to jelly).

Improved inventory management: By understanding buying patterns, retailers can predict demand and stock shelves accordingly, reducing stockouts and overstocking. Simplicity and interpretability: The rules generated by Apriori and FP-Growth are easy to understand, allowing for clear decision-making.

CNN-BiLSTM (Deep Learning):

This approach utilizes deep neural networks to analyze vast amounts of customer data. Its main contributions include:

Handling complex relationships: Deep learning models can capture intricate relationships between items that go beyond simple co-occurrence. This allows for identifying more nuanced buying patterns.

Scalability and adaptability: These models can handle large datasets and adapt to changing customer behavior over time.

Potential for personalization: Deep learning can be used to personalize recommendations for individual customers based on their past purchases and similar customer profiles.

Complexity: Deep learning models are more complex to implement and require significant computational resources.

Interpretability: The underlying logic behind a deep learning models can be less transparent compared to Apriori and FP-Growth rules.

2.6 Conclusions

Our market basket analysis project has been a success! By leveraging data mining techniques, we were able to uncover valuable insights into customer purchasing behavior. Analyzing historical transaction data allowed us to identify frequently bought-together items (frequent itemsets) and the relationships between them (association rules). This newfound knowledge empowers businesses to make informed decisions that can significantly improve customer experience and boost sales.

Key Achievements

One of the major achievements of this project was uncovering hidden patterns in customer purchases. These patterns, which might not have been readily apparent before, provide businesses with a deeper understanding of how customers navigate product categories and what influences their buying decisions.

This knowledge translates into practical applications. By identifying strong associations between products, businesses can optimize store layouts and online product recommendations. Placing frequently bought-together items in close proximity, whether physically in a store or virtually on a website, can encourage impulse purchases and ultimately increase basket size.

Market basket analysis also empowers businesses to develop targeted promotions and marketing campaigns. The insights gleaned from the analysis can be used to offer discounts or bundles on frequently purchased items. Additionally, businesses can personalize product recommendations and tailor marketing messages based on specific customer buying habits.

Finally, understanding customer preferences through market basket analysis allows businesses to optimize inventory management. By stocking up on frequently bought-together items and anticipating demand fluctuations, businesses can reduce stockouts and improve inventory turnover.

Impact and Benefits

The insights derived from market basket analysis can have a significant impact on a business's bottom line. Strategic product placement and targeted promotions can encourage customers to buy more, leading to increased sales. Additionally, a well- organized store layout, personalized recommendations, and efficient inventory management all contribute to a more positive customer experience, ultimately leading to improved customer satisfaction.

Perhaps the most important benefit of market basket analysis is that it provides data-backed evidence to support strategic decision-making. This allows businesses to make informed choices that give them a competitive advantage in the marketplace.

CHAPTER 3
CONCLUSIONS AND FUTURE SCOPE

CHAPTER: 3 CONCLUSION AND FUTURE SCOPE

3.1 Conclusion

General Discussion

Market basket analysis is a very useful technique for finding out co-occurring items in consumers shopping baskets. Such information can be used as a basis for decisions about marketing activity such as promotional support, inventory control and cross-sale campaigns. Tracking not so apparent product affinities and leveraging on them is often seen as a real challenge in the retail business. Even though most of the generated rules are somewhat predictable for a cosmetic store, they still provide value to the retailer. The problem with trivial rules is often found in the marketing literature, but solely depends on the size and type of store. In this research, the stores of the cosmetic chain are rather small ones and the number of transactions is not as big as the number in a big hypermarket, for example. Moreover, the assortment is somewhat limited due to the fact that the stores represent and sell mainly products of a certain cosmetic company. Therefore, it is a bit difficult to mine unusual and interesting rules. However, it is important for the retailer to know exactly which products are purchased together and in what time of the year. The generated rules may not be unusual and interesting, but they are useful and actionable.

Academic contribution

The market basket problem can be seen as the best example of mining association rules. Discovering association rules has been a well-studied area for the past decade. Building up on previous researches by using established methods for mining association rules allowed for discovering useful information for the retailer. After aggregating the data and finding product affinities, the multinomial logistic regression extends the analysis by adding up some probabilities of a consumer purchasing certain products in different seasons and in certain times of the day. Evaluating probabilities of a category membership depending on the two factors – season and time of the day provides the retailer with better understanding of consumers’ needs and suggests action for advertising.

Overall overview of contributions

The contributions of this project are as follows:

1. Products purchased in bundles of 2 and 3 were found for all the four stores of the cosmetic chain.
2. Association rules were generated with the supporting probabilities and importance.
3. Dependency networks are used to visually represent the product interrelationships.
4. Average values per sale and overall value of bundle were estimated for every store of the chain (see Appendix B).
5. The multinomial logistic regression provides a model for predicting the likelihood of a consumer purchasing an item from a certain product category at a specific time of the day and in a specific season.
6. The multinomial logit also compares the likelihood of choosing a product out of several options.

Managerial Implications

In the recent years, more and more retailers are seeking competitive edge through advanced and innovative technology. Market basket analysis is the next step in the retail evolution. Applications of association rule mining are growing rapidly in different sectors – from analysing debit and credit card purchases to fraud detections. Mining into big data provides managers with a unique window into what is happening with ones business so that they can implement strategies efficiently. Obscure patterns can be discovered using market basket analysis which can help for planning more effective marketing efforts. It can be used not only for cross-sale and up-sale campaigns, but for managing better inventory control and satisfying shoppers' needs. Almost all departments of a company can benefit from a single analysis – not only the high levels of Management but also Store operations, Merchandising and Advertising and Promotion departments.

3.2 Future Scope

Our market basket analysis project has successfully combined traditional methods (Apriori & FP-Growth) with a cutting-edge CNN-BiLSTM model. This approach has provided valuable insights into customer buying patterns, revealing the "why" behind frequently bought-together items through traditional methods and leveraging the deep learning model's ability to capture more nuanced relationships. However, the potential for even deeper customer understanding remains.

Looking ahead, we can explore advanced deep learning architectures like transformers and graph neural networks (GNNs). Transformers excel at capturing complex buying patterns, while GNNs can effectively model relationships between items themselves. Additionally, incorporating external data sources like weather data, social media trends, and promotional campaigns can enrich our analysis and provide context for predictions. Imagine a model that predicts a surge in sunscreen purchases based on sunny weather forecasts or tailors recommendations based on trending social media interests.

By moving towards real-time personalization, we can create a dynamic shopping experience. Integrating the model with a recommender system could allow for suggesting items based on a customer's current basket, browsing history, and even external factors like time of day. This future-focused approach holds immense potential for unlocking further customer insights, driving targeted promotions, and ultimately boosting sales and customer satisfaction.

BIBILIOGRAPHY

[1] M. Kaur and S. Kang, “Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining,” Procedia Comput. Sci., vol. 85, no. Cms, pp. 78–85, 2016.

[2] A. Mansur and T. Kuncoro, “Product Inventory Predictions at Small Medium Enterprise Using Market Basket Analysis Approach-Neural Networks,” Procedia Econ. Financ., vol. 4, no. Icsmed, pp. 312–320, 2012.

[3] X. Su, “Intertemporal Pricing with Strategic Customer Behavior,” Manage. Sci., vol. 53, no. 5, pp. 726–741, 2007.

[4] G. Armstrong, S. Adam, S. Denize, and P. Kotler, Armstrong, G., Adam, S., Denize, S., & Kotler, P. Pearson Australia., 2014.

[5] E. Sherman, A. Mathur, and R. B. Smith, “Store Environment and Consumer Purchase Behavior: Mediating Role of Consumer Emotions,” Psychol. Mark., vol. 14, no. 4, pp. 361–378, 1997.

[6] N. Jothi, N. A. Rashid, and W. Husain, “Data Mining in Healthcare - A Review,” Procedia Comput. Sci., vol. 72, pp. 306–313, 2015.

[7] A. Bertoni and T. Larsson, “ScienceDirect Data Mining in Product Service Systems Design: Literature Review and Research Questions,” Procedia CIRP, vol. 64, pp. 306– 311, 2017

[8] V. a. I. K. Umayaparvathi, "Automated feature selection and churn prediction using deep learning models," International Research Journal of Engineering and Technology

[9] P. a. D. A. Ghadekar, "Image-Based Product Recommendations Using Market Basket Analysis," in 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), IEEE, 2019

[10] A. Sharma, "Retail Customer and Market Proclivity Assessment using Historical data and Social Media Analytics," 2020

[11] Rakesh Agrawal, Tomasz Imielinski, Arun Swami "Mining Association Rules between Sets of Items in Large Databases", 1993

[12] Market Basket Analysis Using Apriori algorithm "İstanbul, 2018"

[13] Parallel Data Mining for Association Rules on Shared-Memory Multi-Processors "M.J. Zaki, M. Ogihara, S. Parthasarathy, and W. Li May 1996"

[14] Set-oriented data mining in relational databases (M. Houtsma and A. Swami, 1993)

[15] Analyzing Association Rule Mining and Clustering on Sales Day Data with XLMiner and Weka A. M. Khattak, A. M. Khan, Sungyoung Lee*, and Young-Koo Lee 2010

[16] Variable Selection for Market Basket Analysis Katrin Dippold Harald Hruschka, February 2010

[17] I. a. R. M. a. H. M. a. N. N. a. L. F. a. N. F. Fauziah, "Market Basket Analysis with Equivalence Class Transformation Algorithm (ECLAT) For Inventory Management Using Economic Order Quantity (EOQ)," 2022.

[18] S. C. a. L. J. a. Y. N. Chintala, "Browsing the Aisles or Browsing the App? How Online Grocery Shopping Is Changing What We Buy," How Online Grocery Shopping is Changing What We Buy, 2022.

[19] Wang, Zhanpeng, et al, "Optimal retail sales strategies for old and new products in monopoly and horizontal competition scenarios," Journal of Retailing and Consumer Services, vol. 71, p. 103218, 2023.