Market Basket Analysis & Recommendation System Using Association Rules

Thesis · August 2020

DOI: 10.13140/RG.2.2.16572.05767

1 author:

Shruthi Gurudath
Griffith College Dublin
2 PUBLICATIONS **3** CITATIONS

SEE PROFILE

Market Basket Analysis & Recommendation System Using Association Rules

Shruthi Gurudath

2990078

Submitted in partial fulfillment for the degree of

Master of Science in Big data management and Analytics

Griffith College Dublin

June,2020

Under the supervision of

Osama Abushama

Disclaimer

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the Degree of Master of Science in in Big data management and Analytics at Griffith College Dublin, is entirely my own work and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfilment of the requirements of that stated above.

Signed: Shruthi G **Date: 12/06/2020**

Shruthi Gurudath
2990078

Acknowledgements

This project would not have been possible with out the support of my family, professors and my friends.

Osama Abushama: My supervisor for this project, has given so many valuable suggestions on implementations and correcting me on cases where I had deviated from the project. He had always given me a helping hand when I was confused or and pushed me in the right direction.

Table of Contents

Acknowledgements iii
List of Figures 6
ABSTRACT 7
1 INTRODUCTION 8
1.1 Market Basket Analysis 8
1.2 Goals of Project 9
1.3 Overview of Approach 9
1.4 Document Structure 10
2 BACKGROUND 11
2.1 Literature Review 11
2.1.1 Market Basket Analysis 11
2.1.2 Association Rules 11
2.1.3 Recommender System 13
2.2 Related Work 18
3 METHODOLGY 20
3.1 Dataset Collection 23
3.2 Understanding Dataset 26
3.3 Pre-processing Data set 27
3.4 Exploratory Data Analysis 28
4 SYSTEM DESIGN AND SPECIFICATIONS 33
4.1 Apriori Algorithm 35
4.2 Frequent Pattern Growth Algorithm 37
4.3 Recommender System 39
5. IMPLEMENTATION 42
5.1. Apriori Algorithm & FP-Growth Algorithm To Mine Association Rules 42
5.2. Development of Market Basket Recommendation Web application 46
6. EVALUATION 48
6.1. Support of the Frequent Itemsets: 48
6.2. Support and Confidence of Association Rules 48
6.3. Lift value of Associated Rules 49
6.4. Network graph 50

7. ANALYSIS 52
7.1. Analysis on Apriori algorithm 52
7.2. Analysis on FP-Growth algorithm 52
7.3. Comparison between Apriori Algorithm & FP-Growth Algorithm 53
8. CONCLUSION 54
9. FUTURE WORK 55
10. BIBLIOGRAPHY 56

List of Figures

Equation 2: To find Support of an item 15
Equation 3: To find confidence between two items 15
Equation 4 : To find Lift for the effectiveness of rule 16
Equation 5: To determine rules out of n items 17

Figure 1 : Work Flow 21
Figure 2 : Sales Distribution Over Days Of A Week 28
Figure 3 : Frequency of Hourly Orders 29
Figure 4 : Frequency of Sales From Morning To Night 29
Figure 5: Popular best-selling products 30
Figure 6 : Popular Departments 30
Figure 7: Popular Aisles 31
Figure 8: Department wise reorder ratio 31
Figure 9: Frequency Of Orders per customer 32
Figure 10 : System Architecture 34
Figure 11: Illustration On Apriori Algorithm 36
Figure 12 : Illustration On FP-Growth Algorithm 38
Figure 13 : Frequent Pattern Tree Structure 39
Figure 14 : Support Of Frequent Itemsets 48
Figure 15 : Support & Confidence values 49
Figure 16 : Confidence & Lift Values 50
Figure 17 : Network graph of Associated Rules 51

Table 1 : Schema of Dataset 24
Table 2: Comparison between Apriori Algorithm & FP-Growth Algorithm 53

ABSTRACT

Market Basket analysis is a technique applied by retailers to understand customer's shopping behaviour from their stores. The result of the effective analysis may improve supplier's profitability, quality of service and customer satisfaction. Instacart is a company that operates as a fast grocery delivery service in America. The purpose of this project is to make use of anonymized data on customers' transactional orders to focus on descriptive analysis on the customer purchase patterns, items which are bought together and units that are highly purchased from the store to facilitate reordering and maintaining adequate product stock. It can be done by analysing the available data in such way that frequent item set can be found and can be analysed to define an association rule. One of the algorithms which helps in finding association rule for frequent item set and to identify the correlation is Apriori algorithm. The model of the apriori algorithm is developed to explore approaches for the application of the rules of association to recommender system. Minimum confidence and minimum support values used for mining rules are parameters of the foremost existence.

Keywords: Market Basket Analysis, Affinity Analysis, Apriori algorithm, Recommender system

1 INTRODUCTION

1.1 Market Basket Analysis

Market Basket Analysis is a key method known and utilized by substantial retailers to reveal relationships between products, like bread, butter, etc. It works by searching for a mix of products that happen together every now and then in exchanges. To give it another perspective, it enables retailers to recognize connections between things that individuals purchase. With the continuous growth of information technology, massive amounts of data are collected and stored by enterprises. It is very important for enterprises to transform this data into useful information and knowledge for decision making in dynamic markets. This value added information discovered from Market Basket Analysis can be used to support decision making.

If it is known that customers who purchase one product are likely to purchase another product, it is possible for retailers to market these products together, or to make the purchasers of a target prospects for the second product. If customers who purchase diapers are likely to purchase beer, they will be more likely to if beer is displayed just beside a diaper aisle. Though the – young fathers result does make sense, stocking up on supplies for themselves and for children before the weekend starts is something that someone would normally think of right away. The strength of those relationships is valuable information and can be used to cross-sell or up-sell.

Data is provided by Instacart open source. The Instacart shopping process is as follows. First, an user places their grocery order through the application. Then, a locally crowdsourced shopper is notified of the order, goes to a nearby store, buys the groceries, and delivers them to the user. Dataset details on 3 million orders from more than 200,000 users. Dataset has variables which focused on the orders as well as the time of the orders. So, order related and time related features were created in order to predict whether a product will be reordered or not.

1.2 Goals of Project

The main objective of the project is to make Instacart retailers to understand the current customer's behaviour and to predict future customers' purchasing behaviour. Leveraging customer transaction data can help in understanding customers' purchasing behaviour, offering right bundles and promotions, assortment planning and inventory management to retain customers, improve sales and extend their relationship with customers.

The specific objectives of the project are as listed below

1. To understand the purchasing pattern of products that comprise the customers' basket.

2. To study about many products usually purchased by the customers.

3. To study the most likely products purchased by the customers along with a particular product category.

4. To recommend and suggest products to individual customers.

1.3 Overview of Approach

The present technological time that we live in has made it feasible for business organisations to accumulate extensive data. Currently, Database technology innovation has sufficiently grown to keep these information stacks solid, however, it is significant not to simply keep that information, yet to assess the information to increase the value of the organisation. In today's customer-centred markets, business needs to establish adequate and low advertising techniques that can react to changes in customer perceptions and demands for products. It might also assist business to recognise a whole new market strategy that can effectively target. All together for making key choices on the market strategy, stable, as much as could be and secured proof-based data is needed. With innovation, Data Mining has gotten perhaps the best response to this requirement. Data Mining is the process of refining important data from enormous Databases which includes a tremendous assortment of statistical and computational methods, neural network analysis, clustering, classification and summing up information. The computation of association rules, which is one of the Data Mining techniques implemented by Market Basket Research, is part of this project. The analysis is carried out on the Grocery stores' transaction data for the customers. The research goal is to consider the category of product that is likely to be marketed in conjunction by implementing Apriori and FP-Growth algorithms.

1.4 Document Structure

The document consists of seven chapters. Chapter One provides a brief introduction to the market basket analysis, project objectives, project overview and technology lists. In Chapter Two, addresses a literature review, results of similar research works and the related work has been discussed which provided insights to work on my project. The methodology and high-level design of the proposed system are defined in Chapter Three. The chapter Four, brief the architecture of the system and specifications, including the hardware and software used. The implementation details for the project are provided in Chapter Five. Details about working model, such as evaluation methods and results are provided in Chapter Six. Finally, Chapter 7 outlines findings and work for the future.

2 BACKGROUND

2.1 Literature Review

In recent decades, data mining have played a key role in marketing literature. Market basket analysis is one of the oldest data mining areas, and the best example of mining association rules. Researchers have developed several algorithms for the Rule Mining Association to help users achieve their goals.

2.1.1 Market Basket Analysis

Agrawal, Imieliński and Swami (1993) (Agrawal, n.d.) apparently first used Market Basket Analysis who had a large collection of consumer transaction data previously collected and the association rules between items purchased were discovered. The method was rapidly implemented as a standard method for a number of practical applications in the field of marketing(Chen *et al.*, 2005).

Kanagawa, Matsumoto, Koike and Imamura posted surveys completing open-choice checklists with respect to related allergenic foods(Kanagawa *et al.*, 2009). As a result, researchers found that certain food allergens appear to happen in the same person together. (2017)

Russell et al . (1999) (Singh and Sinwar, 2017)indicated that it was conceivable to use Market Basket Analysis by marketing researchers in developing multi-category decision-making, theoretical model purchasing decisions involving products. Researchers Y.-L. Chen, Tang, Shen, & Hu, in 2005 have proposed to establish inductive hypotheses from theoretical point of view(Chen *et al.*, 2005). The findings indicate the possibility that customers may have mental templates among several other collections of objects, include connections that are complementary in terms of its activities and interests (for example, running shoes and water bottles). At a logical point of view, the findings of this study can be used to make decisions such as displaying the two products close to each other, enhancing the probability that consumers can find and buy easily two goods, rather than just one.

2.1.2 Association Rules

Apriori algorithm was proposed by Ramakrishnan Srikant and Rakesh Agrawal(Agrawal, n.d.), which is one of the most traditional algorithms to find frequent patterns of the Boolean rules. In large relational tables the authors elaborate on the concept of quantitative rules in mining.

Julander(2007), analysed the percentage of customers buying a certain product and the percentage of overall revenue produced by this same product. By creating such associations one can easily find out about the leading products and what their sales share is. It is extremely important to measure which products are the leading products, as a great number of customers come into contact each day with these specific products. Given the large traffic created by departments with leading products, it is essential to use this data to position other similar products in the vicinity. Thus the process of generating association rules for the products has been analysed in the results.

In the research, Raeder and Chawla (2011) followed a different approach to data on mining processes(Aulakh, 2015). They have found exceptional communities (clusters) in data by modelling data as a product network. In the network method, they have shown that inferences between products can be extracted specifically and that the need for a set of aggregate rules is reduced. First, they analysed the characteristics of the successful networks and showed in the results above the defining groups in those networks would expose significant links between rules of combination.

Berry and Linoff(2004) aimed to discover patterns by extracting associations or co-occurrences from the transactional information provided by a store(Berry and Linoff, 2004). Customers who buy bread also often buy several bread related products such as milk, butter, or jam. The results are shown on how it makes sense to place these area units of groups side by side in a retail centre so that customers can quickly access them. Additionally, such related product groups should be placed side-by - side to remind customers of related products and very logically guide them through the centre.

Ibrahim Cil (2012) (Anon, n.d.) has provided a new plan for supermarket-placement problems with association rules and multidimensional scaling evaluation. He took customer receipt and product barcode details as data in his research for the Migros Türk supermarket, which holds an important position in the retail sector, and analysed it using the Apriori algorithm according to the association rules process. Researcher then suggested a new development design for the store based on the resulting rules.

Erpolat(2012) provided information on the Apriori and FP-Growth algorithms which have been widely used in the research of association rules, implemented these algorithms to the consumer shopping data of an approved automotive service and compared the results(Mostafa, 2015). As a result of the analysis, researcher noted that the FP-Growth algorithm is more suitable than the Apriori algorithm to analyse a data set because of the computation speeds.

Kaur and Kang(2016), on dynamic data, suggested an algorithm which would conduct association rule mining(Singh and Sinwar, 2017). They did periodic mining by working with the principle of change modelling. The results are noted on methods to find the outliers and about predicting the future association rules.

The words shopping basket were proposed by Cachon and Kok (2007). The shopping basket establishes the set of products the consumer would like to purchase on a single journey to a shop(Anon, n.d.). According to researchers, the customer would not be treated on subsequent shopping trips if he cannot bring out his shopping cart in a specific store. So, even though the merges for profit are smaller than other products, some products must be offered to the consumer by the retailer. Thus, researchers created a method and approach that tests the productivity of the shopping basket, contributing to the retailer's higher projected return than that of category management.

2.1.3 Recommender System

Recommendation systems are kinds of systems for filtering information which analyse user behaviour data from past times and try to predict the user's preference for items. They focus mainly on individuals without personal experience and are not a differentiated group of customers. Different recommender systems have been studied in multiple areas since the 1990s, such as movies , music, books , articles, social media, and products in general.

Burke outlines five classes of recommendations; content-based, collaborative filtering (CF), demographical, knowledge-based and hybrid systems.

The association rules of mining in the stock market sector were introduced by Paranjape-Voditel and Deshpande(2011). They built a method of portfolio recommendations with analysis of market baskets(Chen *et al.*, 2005). In the results, they have shown how rules are generated after periodic intervals and also provided about related negative loss generating stocks may be substituted with increased stock correlation.

In a pharmacy warehouse, Yazgan and Kusakci(2013) focused on enhancing order collection. The order stacking strategy was suggested for solving this problem where the items to be processed are low in size(Cil, 2012). In order to match the same stack of order receipts with similarity in different pharmaceutical warehouses, stacks were created with genetic algorithm by setting the rules of association between customer orders. For more than one order set in one round of the order picker, there was a reduction in the number of rounds which saved time.

Market Basket Analysis

Market basket analysis determines customer's purchasing patterns by finding significant relationship among the items which they select in their shopping carts. Market Basket Analysis aids the procedure as well as expands target strategy in numerous business associations.

Association Rule

Association rule is one of the most important Data Mining techniques used in Market Basket Analysis. All fruits are sorted in the same aisle in a Super Market, all dairy products are placed together under another aisle. Hence spending time and intentionally investing resources to place the most necessary items in an organised way not only reduces a shopping time of customers, but also helps customers to purchase the most appropriate items one might be keen in clubbing in their Market Basket. Association rule is related to the statement of "what goes with what". The purchase of products by customers at Super Market are termed as 'Transactions'. The magnitude of an associative rule can be derived in the existence of three parameters, namely support, confidence and lift(Kanagawa *et al.*, 2009).

Item Sets

Item sets is the collection of all items in a market basket data, I = { i**1**, i**2**…in }.

Transaction is the group of all transactions, T= { t1, t2..tn }

Every transaction is a specific one and makes up a collection of items from item I. When there are n items in an itemset it is called n- itemset. For example, it is called as 3-itemset {Oranges, Apple, Bread}. If an itemset doesn't have an item, it is called the set as null (empty). Presenting the itemset with more than one item significantly expands the chances of the rules to be listed.

Support Count

The width of the transaction is measured by the number of items contained in a transaction. Support Count, which refers to the number of transactions involving a particular itemset, is an important component of an itemset.

Support of an item or set of item is that the fraction of transactions in our data set that contains number of that particular item product to total number of transactions. Support gives an idea of how many times an itemset has occurred in the overall transactions. For example, in a retail shop, if we consider 100 customers had visited to purchase products. It was seen that out of

100 customers, 50 of them purchased Product A, 40 of them purchased Product B and 25 of them purchased both Product A and Product B. Support of Product A is 50%, Support for Product B is 40% and Support of Product A and B is 25% .Value of Support helps in considering the rules which are worth for further analysis on correlation of a products with other existing products in the store. For instance, if we want to note the item sets which occur at least 50 times in 10,000 transactions, then support = 0.005.With low support value, we will not have enough information on how the products are related to each other and thus helps to find "hidden" relationships.

$$Support\ A = \frac{(Number\ of\ Transaction\ that\ Contains\ A)}{(\ Total\ Transaction)}$$

Equation 1: To find Support of an item

Confidence

Confidence is a measure of the likelihood that customer buy product A will buy product B as well. A rule of association is therefore a remark of the form (item set A) ⇒ (item set B) where A is the precedent and B is the consequence. Confidence gives the probability of Consequence occurring on the cart provided with pre-existing antecedents. For frequently appearing Consequent, it doesn't matter what the customer have it in the Antecedent. The confidence of an Association rule, which results very often, will always be of great value. For example,

$$Confidence(A => B) = P(A \mid B)$$

$$= \frac{(\ Number\ of\ Transaction\ that\ Contains\ A\ and\ B\ )}{(Total\ Transaction\ that\ Contains\ A)}$$

Equation 2: To find confidence between two items

Of the 50 customers who purchased Product A, 25 have purchased Product B. It ensures if somebody buys product A, they could buy product B by the probability of 50%.

Lift

In comparison to a random selection of a transaction, the ratio shows how effective the rule is in finding consequences. In general, lifting is higher than one implies that the rule has some usefulness. A lift greater than one indicates that the presence of A has increased the probability of generating B in this transaction. The role of A has decreases the chances the role of product B occurrence if lift value is smaller than one.

$$\frac{Lift\ (A => B) = \ (Confidence(A => B))}{(Support(B)\ )}$$

Equation 3 : To find Lift for the effectiveness of rule

A lift value of 1.25 implies that chance of purchasing product B would increase by 25%.

Fundamental rules of association grant the phenomenon of one item and the inclusion of another. Through the use of Data Analytics, the process of trying to uncover association rules includes the following steps:

Step 1 : Set up the data in the transaction presentation. An association algorithm demands input information to be arranged in transaction format t*x* = { ***i1,i2,i3*** }

Step 2 : Short-list collection of items that often occur. Item sets are object aggregation. An association algorithm considers the most commonly occurred items and excludes the least occurred items, making increasingly relevant the ultimate rule that will be pulled to the next level.

Step 3 : Generate the association rules applicable from the item sets. Ultimately the algorithm produces and filters rules based on the measurements being tuned.

Frequent Itemset Generation

For an association analysis of **n** items, 2^n -1 item sets can be discovered apart from Null item sets. As when the items increases, the number of item sets also increases exponentially. Therefore it is necessary to pick and fix a minimum support threshold to erase item sets that takes place less frequently in the transaction world.

Why is frequent extraction of items a difficult issue?

A simple brute-force strategy will be intended to create all reasonable items-sets and compute across transactions for each candidate, counting the number of times the candidate sees if that item-set meets the appropriate support count. For instance , a typical supermarket may have more than 10,000 separate products in stock. The number of items-sets of size 2 is Choose(10,000, 2) = 49,995,000, at which Choose(n, k) is the range of different items ways to choose k items from n items. The number of possible items-sets of size 9 is Choose(10,000, 9) = 2,745,826,321,280,434,929,668,521,390,000, which is a lot.

Extracting rules from frequent item sets

A brute force solution is to determine the rules for the mining association with Support and confidence for each rule.

$$R = 3n – 2^{(n+1)} + 1.$$

Equation 4: To determine rules out of n items

In a dataset of n items, R rules can be found. This process extracts all the rules with confidence higher than a minimum confidence threshold. From generating frequent items and extracting steps helps in producing hundreds of rules even for a dozen of items. Therefore, to filter out less frequent and less appropriate rules in the search space, it is very important to set a fair support and confidence threshold. The rules derived can also be assessed with measures of support, confidence and lift. In significant to computational time requirements, it is much more costly to identify all the frequent item sets above minimum support value than extracting the rules out of frequent item sets. More explicitly, provided the total number of possible rules is neither on the left nor the right side of the rule, R, should be empty.

There are few algorithmic ways to measure the frequent item sets efficiently. The algorithms Apriori and Frequent Pattern (FP)-Growth are two of the most common algorithms for the analysis of associations.

Product Recommendation Using Association Rule

In order to recommend a product, the main aim of data mining is to create a model. The model builders must derive information from historical data and represent it in such a way as to be able to adapt the resulting model to new situations. The data sets analysis process extracts useful information on which to apply one or more data mining techniques to discover

previously unknown patterns within the data, or find trends in the data which can then be used to recommend trends or behaviour patterns. It is human nature to know what the future holds and to advise. The recommendation covers the forecast of future events by using sophisticated methods such as machine learning based on historical data observed previously. Through using different strategies such as sampling, correlating and so on, historical data is obtained and transformed.

Recommendation system can be split into four stages:

1. The collection and pre-processing of raw data;
2. Convert pre-processed data into an easily achievable form using the selected machine learning method such as Apriori or FP Growth algorithm;
3. Create a model of learning (training) using transformed data;
4. Use the previously developed set of association rules to report recommendations to the user;

2.2 Related Work

In the course of Apriori association rules, Shanta Rangaswamy and G. Shobha[5] presented a method by which genetic algorithm(Mostafa, 2015) applies. The proposed system can predict the rules by using genetic algorithms which in the developed rules contain negative attributes along with more than one consequent part of the attribute. The purpose of the strategy was to enforce associated data mining rules utilizing genetic algorithms to boost the effectiveness of obtaining the database-preserved logfiles and improves the complexity by reducing the time required to scan large server-preserved databases.

In the management of product placement in supermarkets, Raorane has found out that Market Basket Analysis could be used. This method can show that the vendor benefits more. The Data Mining tool is thus to improve the strategy of placing the product on the shelf. Market Basket Analysis has been used to analyse the frequent transactions carried out by the customers using customer support and trust when purchasing related items.

Application of data mining in education field was presented by Omprakash Chandrakar. Association rule mining is used in order to evaluate academic achievement in their exams and to predict the outcome of the next exams(Madani, 2009). This forecast allows students and teachers to distinguish subjects that need extra preparation well before the semester begins.

The rules found through the mining of association rules are used to predict the outcome of the next exam. This forecast may be used to help students identify the concepts they need to focus on at the very start of their semester.

In the current project, by applying association rules on Instacart transactional data of the customers it doesn't extract preferences of the individual customer rather it does specifies the connections at product levels of each transactions for all the customers. There by, mined rules are used for the recommendation system. Perhaps, the strong association rules are supportive for the recommender system in the approach of suggestions of most likely corelated with similar products. Large datasets have been used for the analysis of the pattern and in order to overcome the memory and computing issues, analysis of the data has been done in Google Collab with the usage of GPU while unstacking the tractional data of the customers.

3 METHODOLGY

The aim of a recommending system is to produce meaningful recommendations for items or products of interest to a collection of users. The fundamental algorithm such as Apriori and Fp Growth, collects knowledge about the preferences of people and recognizes that when people buy spaghetti and wine, they are often generally interested in gravies. Association rule is the key part in developing a recommendation engine. The Association Rule produces a number of rules after running on a data set with details from past shopping baskets. Each rule includes a product name collection as an antecedent, one product name as a consequence, and a few class measures, such as antecedent support, consequent support, support, confidence and lift.

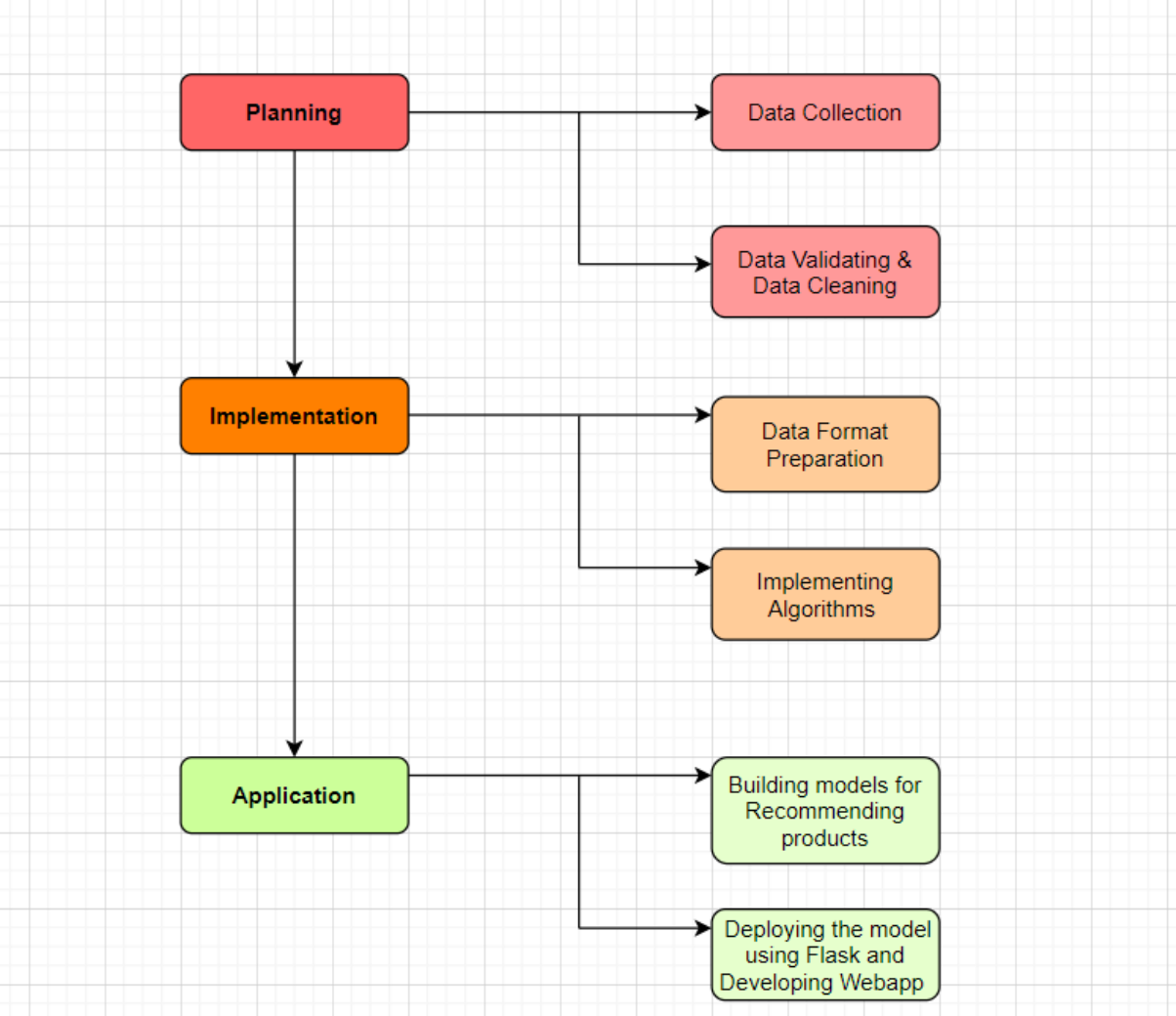

Figure 1 : Work Flow

Machine learning is the study area that helps computers to learn without being specifically programmed. Typically they are being used to overcome various types of life challenges. Now a days, python libraries, frameworks and modules are very simple and powerful as compared with the old days. Python has replaced many of the industry's languages, one of the reasons for this is its vast collection of libraries, and is one of the most popular programming languages for this task today. Python libraries used in the project are:

Pandas

Pandas is an open source, Python licensed library that offers high-performance, easy-to-use data structures, and data analysis tools to the Python programming language. The Data Frame is the core data structure. Data frame allows tabular data to be stored and manipulated in

observation rows and variable columns. A broad variety of stored data types is available, such as CSVs, TSV's (Tab separated values), JSONs (Hypertext Mark-Up Language), and more. Pandas can read different types. A Data Frame consists of both a row and a column index, a two-dimensional set of values. A series is a special collection of index values.

In our project, we have converted dataset CSV files to data frames:

1. order_products_train_df
2. order_products_prior_df
3. orders_df
4. products_df
5. aisles_df
6. departments_df

Numpy

NumPy is an array-processing application for general purposes.
It stands for 'Numerical Python'. It is a library of multidimensional array objects, and a set of array processing routines. NumPy has functions built in for linear algebra and the generation of random numbers.

Matplotlib

Is the art of displaying data through charts, icons, presentations and more. It is most common to translate complex data for a non-technical audience into comprehensible insights. Matplotlib is one of the most powerful Data Visualization Python packages used. This is a cross - platform framework designed to make Two dimensional graphs from records in arrays. This also provides an object-oriented API which helps, for example, to embed plots into implementations using Python GUI toolkits such as PyQt.

Seaborn

Seaborn is an enhancement to matplotlib and not a substitution for it. The reason for this is that it is placed on top of matplotlib and you will often explicitly invoke matplotlib functions to draw simpler plots already available through the namespace pyplot. Matplotlib is completely scalable but it can be difficult to know what settings to change to achieve an appealing plot. Seaborn comes with a number of custom themes to track the matplotlib look and a high level

user interface. It is closely integrated with the PyData stack, including support of SciPy and stats models data structures for NumPy and Pandas, and statistical routines.

MLxtend

MLxtend is a library which implements a range of core machine learning and data mining algorithms and utilities. The primary goal of MLxtend is to create widely used tools to focus solely consistency with existing machine learning libraries on user-friendly and intuitive APIs. While MLxtend enforces a wide range of functions, highlights include sequential selection feature algorithms, stacked generalization implementations for classification and regression, and frequent pattern mining algorithms. MLxtend offers a variety of utilities that draw on Python 's scientific computing stack and increasing its capabilities.

3.1 Dataset Collection

The data set used for the project is a reliable set of files showing the orders of customers over time. They are unidentified user personal details, and contain a sample of over 200,000 users of Instacart and 3 million records of grocery. For an individual customer, Instacart gave anywhere in the range of 4 and 100 of their grocery orders, with the arrangements of products shopped together in each order, hour of day, day of the week, the order was placed in, and a specific measure of time between orders was specified.

We have imported a total of five datasets. Datasets were derived from an open Kaggle competition.

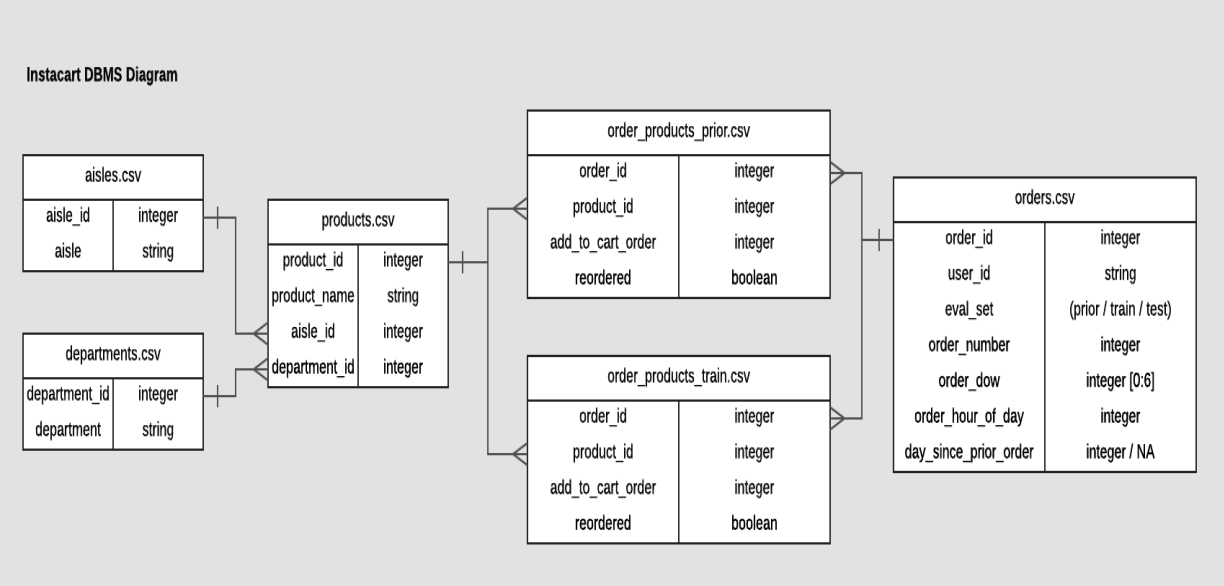

Table 1 : Schema of Dataset

1. **orders (3.4m rows, 206k users):**

This file specifies the order ID, customer ID, the day and hour on which an order was placed, days from the last order column, gives the time difference between two orders and includes NULL value for each user's first order. Eval set indicates whether the order is a prior order, a train or a test. All but every order is classified as prior, except for the last order. Each user's last order is classified as either a train or a test.

- order_id: order ID
- user_id: customer ID
- eval_set: to which evaluation this order belongs
- order_number: The order sequence number (1 = first, n = nth)
- order_dow: order was placed on the day of the week
- order_hour_of_day: the time of day on which the order was placed
- days_since_prior: capped at 30 days from last order (with NAs for order number = 1)

2. **products (50k rows):**

This file contains product identification information, product name, aisle ID, and department ID

- product_id: product identification number

- product_name: name of the product
- aisle_id: foreign key
- department_id: foreign key

3. **aisles (134 rows):**

This file mentions about name of the aisle and aisle id.

- aisle_id: aisle identifier
- aisle: the name of the aisle

4. **departments (21 rows):**

This file gives details on department id and the name of the department.

- department_id: department identifier
- department: the name of the department

5. **order_products (30m+ rows):**

This file provides data on the order where each product was added to the basket and denotes 1 if this product was previously ordered by the user, otherwise 0.

- order_id: foreign key
- product_id: foreign key
- add_to_cart_order: order in which each product was added to cart
- reordered: 1 if this product has been ordered by this user in the past, 0 otherwise

where SET is one of the four following evaluation sets (eval_set in orders):

- "prior": orders prior to that users most recent order (~3.2m orders)
- "train": training data supplied to participants (~131k orders)
- "test": test data reserved for machine learning competitions (~75k orders)

3.2 Understanding Dataset

Aisles

The dataset includes **134 aisles**. Here are a few sample names of aisles, energy granola bars, fast snacks, meat preparation marinades, other fresh salad soups, speciality cheeses.

	aisle_id	aisle
0	1	prepared soups salads
1	2	specialty cheeses
2	3	energy granola bars
3	4	instant foods
4	5	marinades meat preparation

Departments

This dataset includes **21 departments**. All department names are listed in alphabetically ordered below.

```
array(['frozen', 'other', 'bakery', 'produce', 'alcohol', 'international',
       'beverages', 'pets', 'dry goods pasta', 'bulk', 'personal care',
       'meat seafood', 'pantry', 'breakfast', 'canned goods',
       'dairy eggs', 'household', 'babies', 'snacks', 'deli', 'missing'],
      dtype=object)
```

Products

Within 134 aisles and 21 departments, there are **49,688** items in the file.

	product_id	product_name	aisle_id	department_id
0	1	Chocolate Sandwich Cookies	61	19
1	2	All-Seasons Salt	104	13
2	3	Robust Golden Unsweetened Oolong Tea	94	7
3	4	Smart Ones Classic Favorites Mini Rigatoni Wit...	38	1
4	5	Green Chile Anytime Sauce	5	13

Orders

Let's examine a user's construct. For eg, user ID 1 had made 10 orders in advance (order number from 1 to 10), the last order was a train (eval set).Notice that the first command (order number 1) has no meaning for day since prior order, since it is the first order without previous record.

	order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
0	2539329	1	prior	1	2	8	NaN
1	2398795	1	prior	2	3	7	15.0
2	473747	1	prior	3	3	12	21.0
3	2254736	1	prior	4	4	7	29.0
4	431534	1	prior	5	4	15	28.0

3.3 Pre-processing Data set

To work with missing values

To replace nulls with non-null values, a technique known as imputation has been used.

Let's calculate to total number of nulls in each column of our dataset. The first step is to check which cells in our Data Frame are null by using .isnull() and then to count the number of nulls in each column we use an aggregate function for summing .isnull.sum()

Replace numeric values of Day of a week field to textual values

Replacing **days_since_prior_order** field for Null values as **First Order**, since very First Order were left blank in the original dataset.

The days of a week when purchased had been mentioned in the numeric values in the original dataset ,therefore numeric values are replaced with textual values for the ease of understanding.

3.4 Exploratory Data Analysis

What is the frequency of the orders according to days of a week?

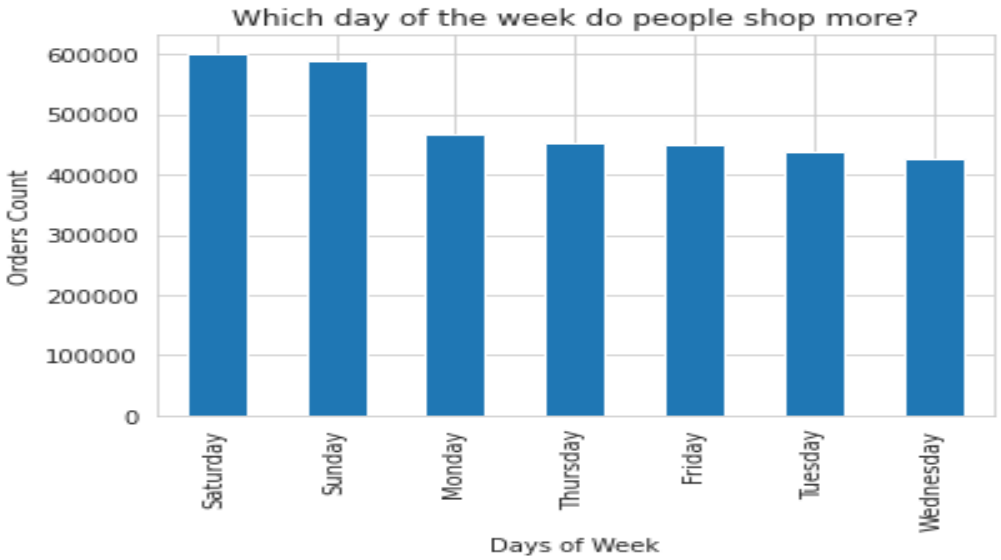

Figure 2 : Sales Distribution Over Days Of A Week

Most sales occur on Saturday and Sunday followed by Monday. On the weekends, consumers could shop for their weekly grocery. The other days of the week will not have a wide difference. The disparity is really not that significant , for example, between Wednesdays and Tuesdays.

When do Instacart customers prefer to shop?

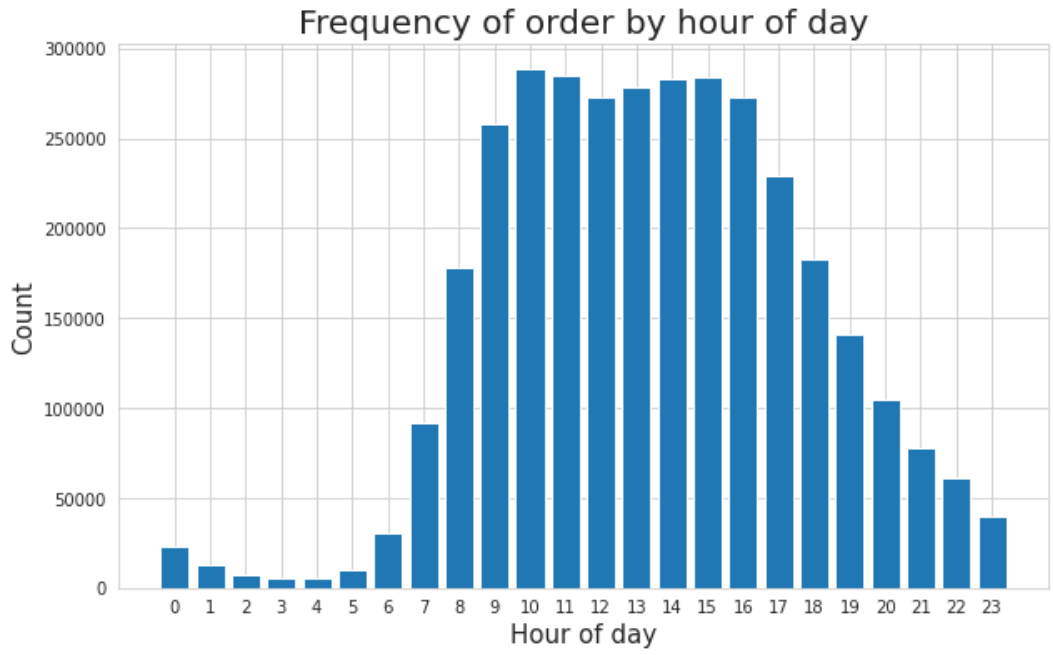

Figure 3 : Frequency of Hourly Orders

Most transactions happen between 10 am and 4 pm.

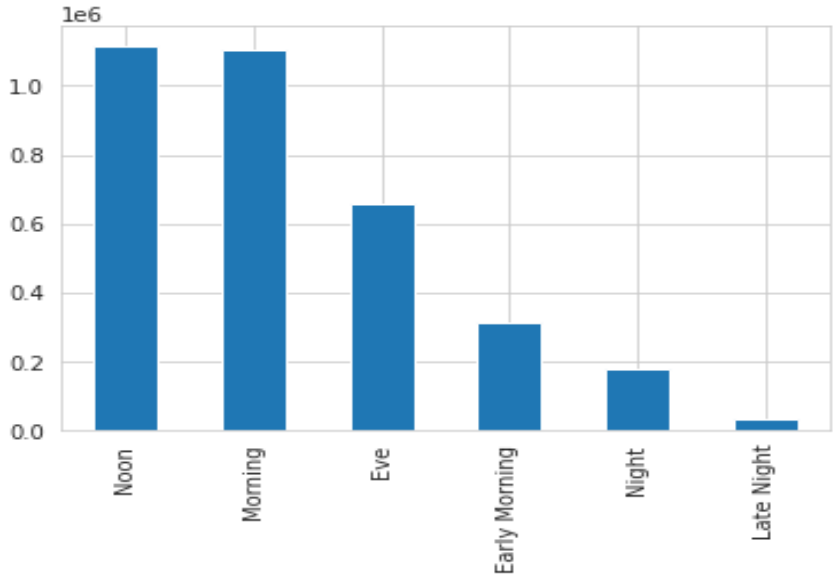

Figure 4 : Frequency of Sales From Morning To Night

From the above graph, we can infer that the most of the orders are occurred in Morning to Noon time. Hence Instacart can accordingly plan to hire persons for delivery during days shifts.

Which are the top 10 best-selling products?

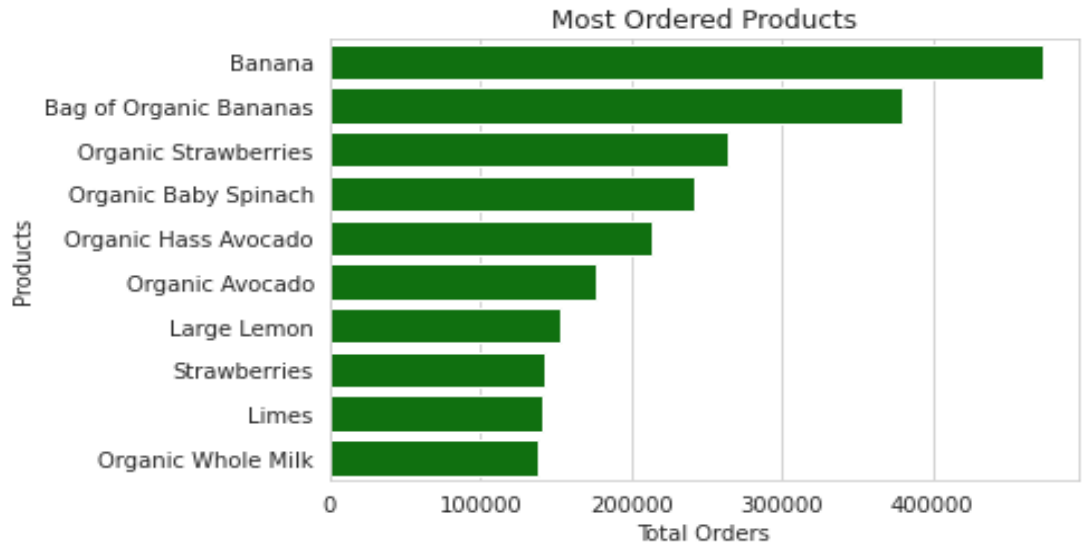

Figure 5: Popular best-selling products

Top 10 most popular selling products are Banana, Bag of Organic Bananas, Organic Strawberries, Organic Baby Spinach and followed by Organic Hass Avocado. We can observe that organic products are most popular in the Instacart stores.

Which Departments are marketed more popularly?

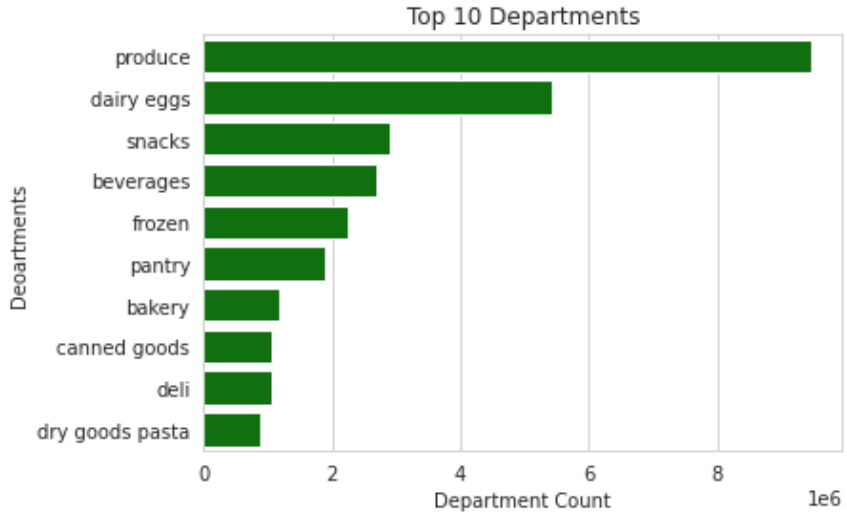

Figure 6 : Popular Departments

Top 10 most popular Departments are produce, dairy eggs ,snacks, beverages and followed by frozen and pantry.

Which Aisles are marketed more popularly?

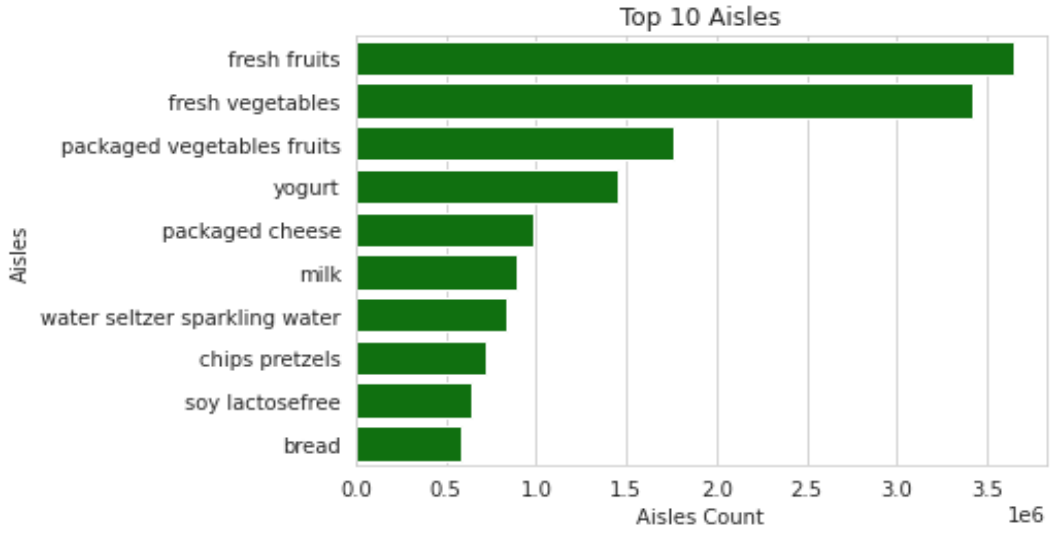

Figure 7: Popular Aisles

Top 10 most popular aisles are fresh fruits, fresh vegetables, packaged vegetables and fruit, followed by yogurt and packaged cheese.

Department wise reorder ratio

Dairy eggs have the highest reorder ratio. The beverages and produce stands in the second place, that also states the very same reorder ratio. Personal care is the department which has the least reorder ratio.

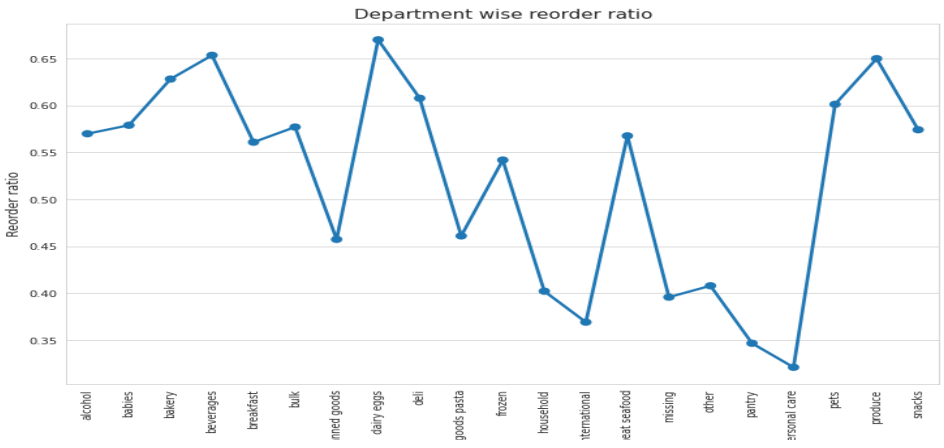

Figure 8: Department wise reorder ratio

What are the minimum and maximum orders received from customers?

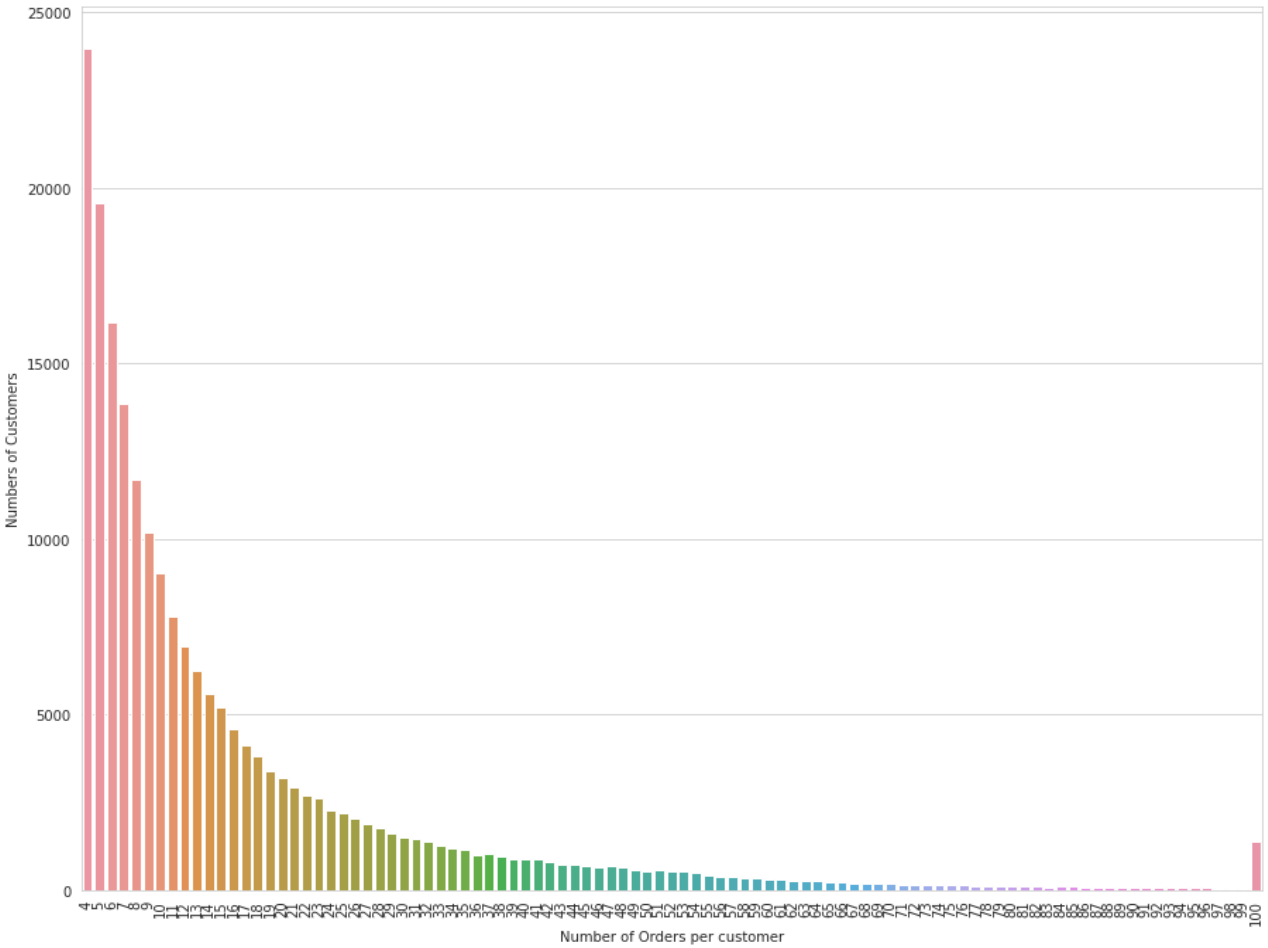

Figure 9: Frequency Of Orders per customer

Here, there are only 4 orders from 23,986 customers, only 5 orders from 19,590 customers. As the number of customers orders increases, the number of customers ordering decreases. Customers large percentage make 4 to 12 orders. When a company figures out a way to increase the number of repeat customer orders, the sales will increase.

4 SYSTEM DESIGN AND SPECIFICATIONS

Technology has introduced new buying behaviours to customers, and the rise of e-commerce has provided new possibilities with retailers to reach those customers. Retailers are increasingly turning to data analytics to identify opportunities for improvement and assist customer product offerings. Market basket analysis has become one of the main techniques major retailers use to discover interconnections between products. It operates by looking for combinations of items which mostly occur together in transactions. It helps retailers to define relationships between the products that customers purchase, to put it another way. Based on the principle of strong rules, association rules are commonly used to analyse retail basket or transaction data and are intended to define strong rules found in transaction data using interesting steps.

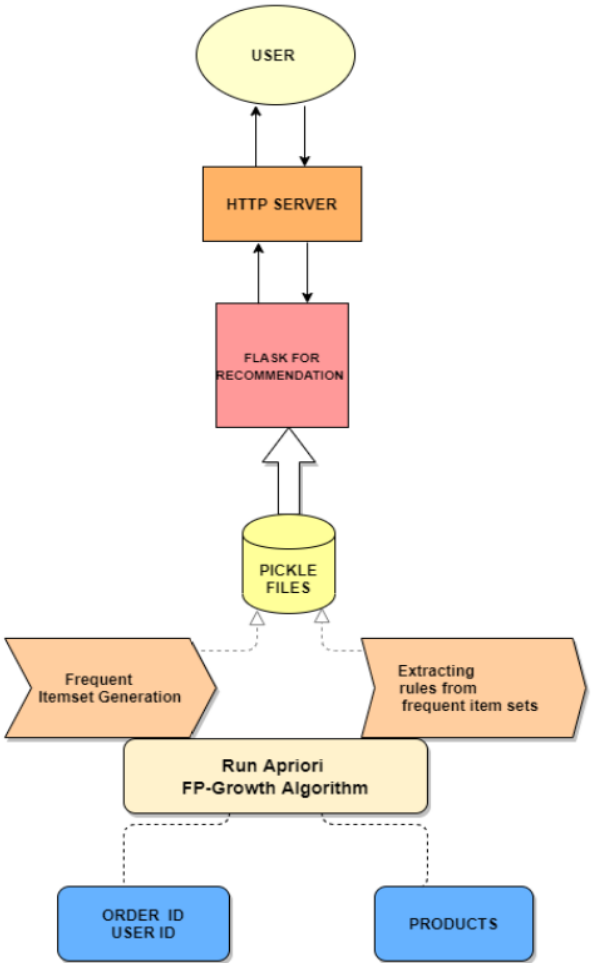

Figure 10 : System Architecture

The principal disadvantages of association rule algorithms are:

• Obtaining rules which are not of interest
• Enormous rules discovered
• underperformance Algorithm

At first we describe the basic ideas of analyzing associations and the algorithms we used to mine the frequent patterns from a large dataset. Firstly, we will specify how to eliminate spurious outcomes.

We need to turn the data from the data frame format into transactions before using any rule mining algorithm so that we have all the products purchased together in single row. For example, the format we need is this:

Here we have insta cart transaction dataset, and the matrix of items being purchased together is shown. By applying Apriori and FP Growth Algorithms we can find the items frequency.

4.1 Apriori Algorithm

The Apriori Algorithm is a basic algorithm proposed by Agrawal & Srikant in 1994 for the determination of the frequent itemset for Boolean association rules. The principles of Apriori state that “if an itemset is frequent, then all its subset items will be frequent”. If the support for the itemset is more than the support level, the itemset is “frequent”. The algorithm is based on the prediction of items, which move from the previous stage on a regular basis. The name derived from the term "prior". Apriori algorithm includes the type of association rules in data mining. The rule that states associations between multiple attributes is often called affinity analysis or market basket analysis.

For understanding Apriori principle, let us consider an instance, if the item set {b, d, e} from the dataset which is a frequent itemset, i.e. its support measure(0.35) is greater than the minimum support measures(0.25), then all of its subsets such as b, d, e,{b ,d},{b, e},{d, e} will also be frequent item sets. As a result, Therefore all subtypes b, d, e have to be regular if {b, d, e} is frequent. On the contrary, if any item sets like {a, b} are uncommon then all the supersets must be even rare. The entire segment containing the {a, b} supersets can be trimmed right away. The pruning method of the linear direction relying on the support measure is called as Support-based pruning. The sort of pruning process is achieved by a major objective of the support measure. This characteristic is also known as the antimonotone property of the support measure.

The Apriori algorithm works in two steps:

Prune and Join:

1.Generate all frequent item sets – A frequent item set is an item set that has transaction support above minimum support.

2.Generate all confident association rules from frequent item sets – A confident association rule is a rule with confidence above minimum confidence.

To apply Apriori algorithm on Instacart dataset, the Apriori class is applied that is imported from the Apyori library.

• k-itemset is the itemset which contains: k element number.

- Lk refers to frequent sets of items with; k items.
- Ck corresponds with frequent sets of candidate items with elements; k.

The Apriori function reduces the number of items to be searched to find frequent sets of items. This algorithm continues to identifying 2-itemsets using 1-itemsets, and 3-itemsets using 2-itemsets in an iterative way. This can be generalized as follows; frequent item sets

(k-1) elements are used to find frequent item candidates for k elements.

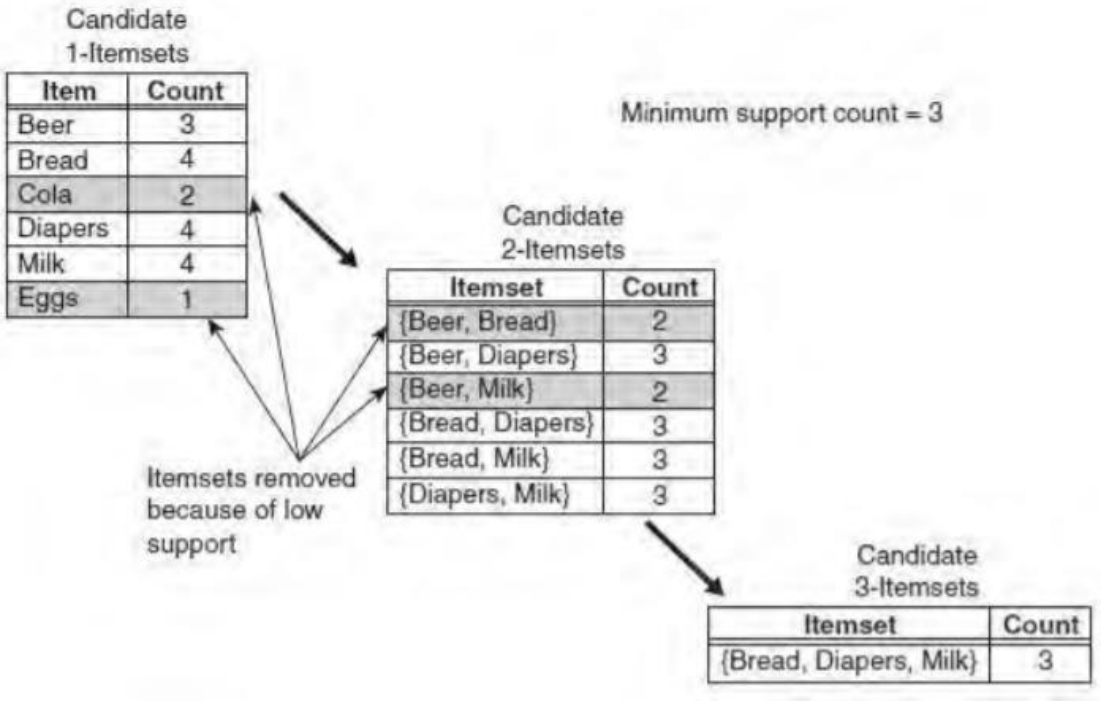

Figure 11: Illustration On Apriori Algorithm

In the above instance, each item is initially considered as a 1-point candidate. Thereafter we count on their support and the itemset appearing less than minimum support count was discarded. As a consequence {Cola} and {Eggs} are removed. In the following iteration, candidate 2-itemset is generated with the help of frequent 1-itemset, because the Apriori principle ensures that all supersets of the rare 1-itemsets must be rare.

The number of candidate 2-itemsets generated by the algorithm is (4C2) = 6, because there are only four frequent 1-itemsets. The performance of the pruning strategy can be demonstrated by counting candidate generated item sets. A brute -force strategy to list al item sets (up to length 3) as candidates will give in 41 candidates.

(6C1) + (6C2) + (6C3) = 6 + 15 + 20 = 41

Where the theory of Apriori is concerned, the number decreases to 13.

(6C1) + (4C2) + 1 = 6 + 6 + 1 = 13

This shows a reduction of 68 percent even in a simple case, in the number of candidate item sets.

The pseudo code shown in Algorithm for the frequent element set generation part of the Apriori algorithm. Let Ck denote the candidate set of k-item sets and Fk denote the frequent set of k-item sets.

4.2 Frequent Pattern Growth Algorithm

The FP-Growth algorithm offers an alternate means of measuring a frequent item collection using an FP-Tree graphic data structure to compact transaction records. One can think of FP-Tree as turning the datasets into a graph format. Instead of the generation and check method used in the Apriori algorithm, FP-Growth generates the FP-Tree first, and uses this compact tree to produce the regular itemset. The FP-Growth algorithm's efficiency depends as to how much compression can be performed while generating the FP-Tree.

The FP-Growth method transforms the problem of repeating the search of the minors and then combining the suffixes in the discovery of broad specific models. With the use of having slightly repetitive objects as a suffix it provides strong efficacy. This approach decreases search costs significantly.

FP-Tree representation

A FP-tree is a compact data structure representing a collection of tree-shaped records. Every transaction is read out and sorted to an FP-tree path. This will come into force until all transactions are read out. The tree remains compact because the paths overlap by different transactions which have common subsets.

FP-Growth algorithm is the main execution mechanism [43];

1. Initially, scan the database and you can find items equal to and above the threshold value.

2. Support values for specific products are displayed in a size (large to small) order.

3. It then produces a tree with only roots.

4. The database is re-scanned for each sample;

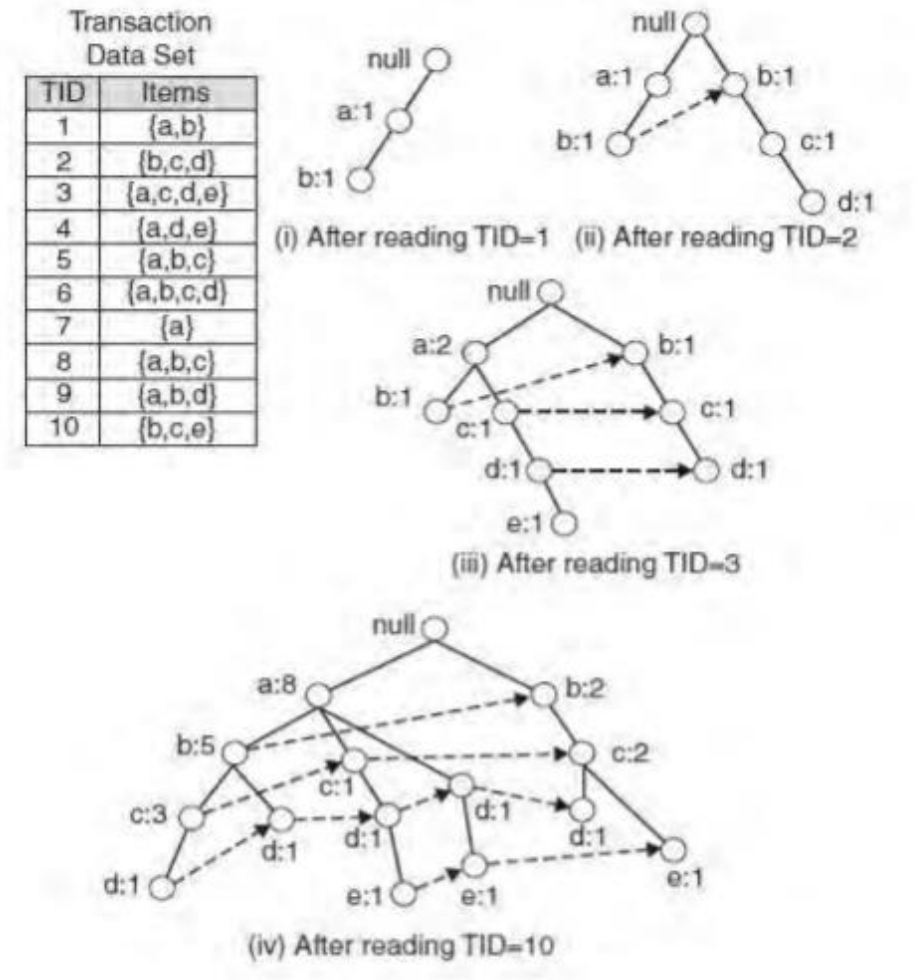

Figure 12 : Illustration On FP-Growth Algorithm

A data collection of 5 transactions and five items can be seen in the diagram. The FP tree structures also appear in the figure after taking the first three transactions. Each node in the tree includes an item's label and also a tracker that represents the number of transactions that have taken the specified path. The way the FP-tree is generated is illustrated below:

1. The data is scanned at first to produce the support value for each item. Items which are not frequent are removed, but at the other side items which are frequent are organized in decreasing order. The figure above shows that a has been the most common item, then c, then d and ultimately e.

2. The algorithm then crosses the data again for the FP-tree structure. The nodes a and b are generated after reading the first transaction {a, b}. The transaction in a tree is then generated from root- > a-> b. Now every node has its count value.
3. Then new nodes are created to represent b, c , and d when the second transaction is crossed {b , c, d}. Then a path is formed by the connection of the b , c and d nodes (root->b->c->d). Whereas the first two transactions involve b, these will not connect because they have a separate predecessor.
4. Then perhaps the third transaction {a, c, d , e} has an initially transacted common predecessor. As long as the item predecessor matches, the path overlaps.
5. The same process goes on until all data is put into the FP-tree.

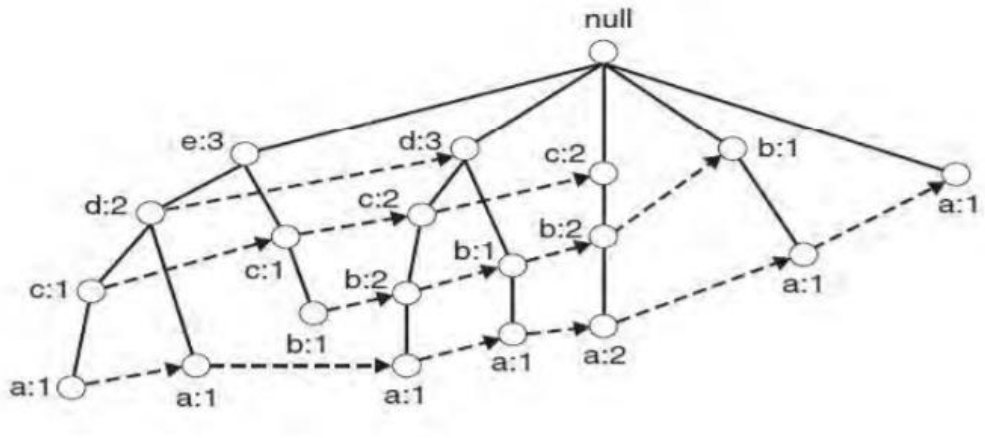

Figure 13 : Frequent Pattern Tree Structure

4.3 Recommender System

An essential part of this thesis research is the association rule mining. The Association Rules for Collaborative Recommenders Framework algorithm is adapted from Apriori. To get an appropriate number of essential rules for the recommendation of the products ,the algorithm needs to be tuned with an appropriate minimum support of the rules during mining.

With only one target user or article at a time, our processing algorithm focuses on mining rules. That has the advantages of:

Since we are only interested in recommending items which a target consumer wants, we need only rules with [antecedent: consequent] in the head rule and we need to mine user associations online, so system efficiency is of great importance.

The rules governing mining associations for recommended systems may be described as follows:

Provided a transaction data collection, a target item, a defined minimum confidence and the number of rules, find in the heads of the rules association rules with the target item such that the rules have the highest support possible and the rules meet the minimum confidence constraint.

We also want specify a high minimum level of confidence and a range for the number of rules before the mining process, instead of a minimum level of support. For the following reasons:

Confidence reflects the degree of similarities between products and help reflects the value of the similarities. Therefore, it is clear that both should be very relevant for making recommendations. The higher the confidence and support, the higher the consistency of the recommendations, we might expect. But before the mining process , it is difficult to choose a proper minimum of confidence and support for each product, since the preferences of customers of products differs. But before the mining process , it is difficult to choose a proper minimum of confidence and support for products, since the preferences of users and the popularities of products differ widely. Probabilities of not getting enough rules for correct recommendation will be high if the minimum level of confidence and support is set too high and in the other way if it tuned to too low value, then likelihoods of getting an utterly unacceptable long runtime to fetch the rules from all the frequent items. Hence small number of rules is good enough for recommendation for each user / article. In view of these factors, the best approach is to define a high minimum level of confidence and a set of rules and to enable the system to find a suitable minimum demand.

In the proposed recommendation system, it only develops association rules with one item in the outcome for the following reasons:

Association rule mining generates usually too many rules even if one confines oneself to rules with only one item in consequence. (so it shoots out the production length by increasing the items) and More complex rules add practically nothing at all to the insights on the data set. Consider the simpler rules that refer to a rule with several items in the outcome, i.e. rules with the same antecedent, but instead with only single items from the effect of the complex rule. All of these rules must necessarily be in the results, since neither their support nor their confidence

can be less than that of the more complex rule. That is, if a rule X Y ->S T , the rules X Y ->S and X Y ->T must also be in the output.

5. IMPLEMENTATION

5.1.Apriori Algorithm & FP-Growth Algorithm to Mine Association Rules

First instinct will be to search for a ready-made algorithm on the scikit-learn. Scikit-learn however does not support this algorithm. Fortunately, Sebastian Raschka's extremely useful MLxtend library has an Apriori and FP-Growth algorithms implement to extract frequent item sets for further analysis. Study of the association is relatively light on the calculation terms and simple to comprehend to non-technicians. Furthermore, this is an unsupervised learning tool that searches for hidden patterns so data analysis and model building are minimal in need. This is a big help for some cases of data exploration and perhaps use certain approaches to pave the way for a deeper dive into the data.

Apriori is a common algorithm for retrieving frequently seen item sets with association rule learning applications. The Apriori algorithm was built to operate on large databases transactions, such as consumer decisions from a store. FP-growth is an improved form of the APRIORI algorithm commonly used for frequent mining patterns. It is used to analyse patterns or associations of data sets in frequent cases.

An itemset is called "frequent" if it meets a level of support defined by the user. For example, if the support value is set at 0.5 (50 percent), a frequent item set is explained as a number of items that occur in at least 50 percent of all transactions in the database totally.

```
]  df_order_items = order_product_merge5[['order_id','product_name']].copy()
   df_order_items.rename(columns={'order_id':'order','product_name':'items'},inplace=True)
   df_order_items['temp']=1
   df_order_items
```

To begin the study of market baskets of Instacart data. Initially, group by the columns to consider to apply the algorithm and also to look at the products from each order ID for the purposes of this analysis.

Reshaping Data frame which contains details on products and order ID using unstack ( ) function

Unstacking, as the name suggests, does exactly the reverse stacking process , it will transform the innermost index of rows back into the innermost index of columns. Unstacking can transfer the data to a shorter but wider data frame (with less rows but more columns).

```
df_order_items = order_product_merge5[['order_id','product_name']].copy()
df_order_items.rename(columns={'order_id':'order','product_name':'items'},inplace=True)
df_order_items['temp']=1
df_order_items
```

	order	items	temp
0	2	Organic Egg Whites	1
1	26	Organic Egg Whites	1
2	120	Organic Egg Whites	1
3	327	Organic Egg Whites	1
4	390	Organic Egg Whites	1
...	...	...	...
31498559	29989	Organic Unsweetened Berry Coconut Granola	1
31498620	29989	Organic Cherry Orchard Fruit Bites Pouches	1
31499010	29991	Island Mango Premium Fruit Snacks	1
31499574	29994	Tomato Cherry On The Vine	1
31499734	29998	Freshly Made. Filled with creamy Ricotta, Aged...	1

284939 rows × 3 columns

With respect to order ID and products that has been purchased is 284939 rows and 3 columns, by applying unstacking we have got 28185 rows × 24423 columns.

```
df_encoder = df_order_items.groupby(['order','items'])['temp'].sum().unstack().fillna(0)
df_encoder
```

items	#2 Coffee Filters	#2 Cone White Coffee Filters	#4 Natural Brown Coffee Filters	& Go! Hazelnut Spread + Pretzel Sticks	0 Calorie Acai Raspberry Water Beverage	0 Calorie Strawberry Dragonfruit Water Beverage	0% Fat Black Cherry Greek Yogurt	0% Fat Blueberry Greek Yogurt
order								
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...
29994	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

28185 rows × 24423 columns

Next, to prepare for performing our mlxtend test, we want to hot encode the data and get 1 transaction per row. We have encrypted our data to show when another product is being sold with it. If there's a null, that means the goods didn't sell together. We want to convert all of the data to become a '1' or a '0' before we start (negative values are reduced to zero, positive values are turned to a 1). This encoding stage can be accomplished by using the following ways:

```
def myencoder(i):
    if i <= 0:
        return 0
    elif i>=1:
        return 1
```

Final step in encoding:

```
df_encoder=df_encoder.applymap(myencoder)
```

API

- **apriori (dataframe , min_support=0.01, use_colnames=False, max_len=None, verbose=0, low_memory=False)**

- **fpgrowth(dataframe, min_support=0.01, use_colnames=False, max_len=None, verbose=0)**

Parameters:

dataframe:

Data Frame pandas Encoded file. Also supports Sparse Data Frames.

min_support: float (Standard: 0.01)

A float for minimum support of the returned frequent item sets between 0 and 1. The support is measured as the transactions where item(s) occur / total transactions fraction.

use_colnames: bool (False by default)

If True, instead of column indices, uses column names of the Data Frames in the returned Data Frame.

max_len: int (None by default)
Maximum length of the generated itemset. If None (default) assesses all feasible item - sets lengths (together under apriori condition).

Verbose: int (Standard: 0)
Represents the list of iterations if > = 1 and low memory is Valid. If = 1 and low memory is False, indicates how many variations there are.
Low_memory: bool (False)
If True, then check for combinations above min support using an iterator. Note further that low memory = True could be used for large dataset when the disk facilities are restricted, since this strategy is approx. 3-6x smaller than average.

Returns
Pandas Column Data Frame ['support', 'item_sets'] of all item_sets >= min support and < than max len (if max len is not None). Every itemset in the 'item_sets' column is frozen set sort, that is a designed-in Python sort that works like sets except that it is immutable.

Frozensets

The items in the column "item_sets" are of the frozenset, which is created-in Python sort equivalent to a Python set but immutable, making it more powerful for certain query or comparison operations. Although frozenset are sets, it doesn't matter what the element order is. I.e., Query

Frequent_Itemsets[Frequent_Itemsets['item_sets'] == {'Banana', 'Strawberry'}]

is equivalent to any of the following three
Frequent_Itemsets[Frequent_Itemsets['item_sets'] == {'Banana', 'Strawberry'}]
Frequent_Itemsets[Frequent_Itemsets['item_sets'] == frozenset(('Banana', 'Strawberry'))]
Frequent_Itemsets[Frequent_Itemsets['item_sets'] == frozenset(('Strawberry', 'Banana'))]

Required to apply the mlxtend 'Apriori' function on Instacart dataset to determine which items are often bought together.

The role 'Apriori' requires to have a minimum support to obtain frequent itemsets. Keeping a support level to large might lead in very few (or even no) outcomes of frequent itemsets and reducing it to low can require a large amount of memory to process data. I initially set the 'min support' to 0.05 for this data but did not receive any results so I altered it to 0.01.

```
freq_itemsets = apriori(df_encoder, min_support=0.01, use_colnames=True)
print(freq_itemsets)
```

The last step is to build the association rules with the function mlxtend 'association rules.' Establish the parameter that is most important (either lift or confidence) to set the minimum confidence level threshold (called min threshold).The 'min threshold' as the level of confidence need back with. We will see only rules with 100% confidence when we set 'min threshold' to 1, for instance.

```
rules = association_rules(freq_itemsets,metric='lift',min_threshold=1)
rules
```

5.2.Development of Market Basket Recommendation Web application

To develop the web application, Flask has been used. Flask is a micro-web application platform that is essentially a collection of tools and libraries that facilitates the creation of web applications using python programming language. The frequent itemset which is greater than the min_support and association rules are saved using pickle library. Pickle is used to serialize and de-serialize a structure of the Python objects. In which object python is transformed to byte stream. The method dump() ,dumps the object into the file stated in the arguments. The method pickle.load() loads the file and stores the deserialized bytes to the model.

- Home page of the Market Basket analysis gives the information on number of products and count of the rules.
- In recommendation page, it allows the user to choose products to cart and displays the added items to the cart along with the list of recommended products with respect to the added items earlier.

- In Exploratory Data analysis page, it displays the graphs on the popular products, departments, aisles and results of the algorithms used for mining the association rules.

6. EVALUATION

6.1.Support of the Frequent Itemsets:

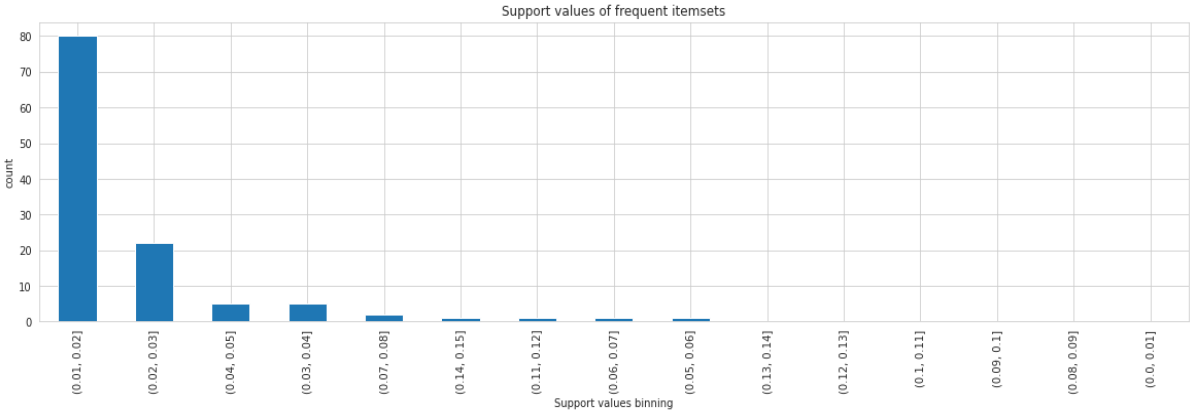

Figure 14 : Support Of Frequent Itemsets

In the above scatter plot ,support of all the frequent itemsets are plotted to evaluate that are greater than min_support value. So we can infer the most commonly used frequent itemsets fall into 0.01 to 0.15 support values.

6.2. Support and Confidence of Association Rules

The rules of association (Pang-Ning et al., 2006) are usually represented in the A / B form, where A (also known as the precedent rule) and B (also known as the consequent rule) are separate and distinct itemsets (i.e. disjoint feature conjunctions). The quality of the rule is usually measured by support for the rule and by confidence. Support for rule is the percentage of data transactions that contain both A and B. This represents the prior chance of (i.e., the frequency observed) in the dataset. Rule confidence is the probability of finding B given A on condition. This describes the implication strength and is given by c(A -> B) = s(A ->B)/s(A). Hence, in the below graph we can see the distribution of support and confidence of the association rules mined from the Instacart frequent itemsets.

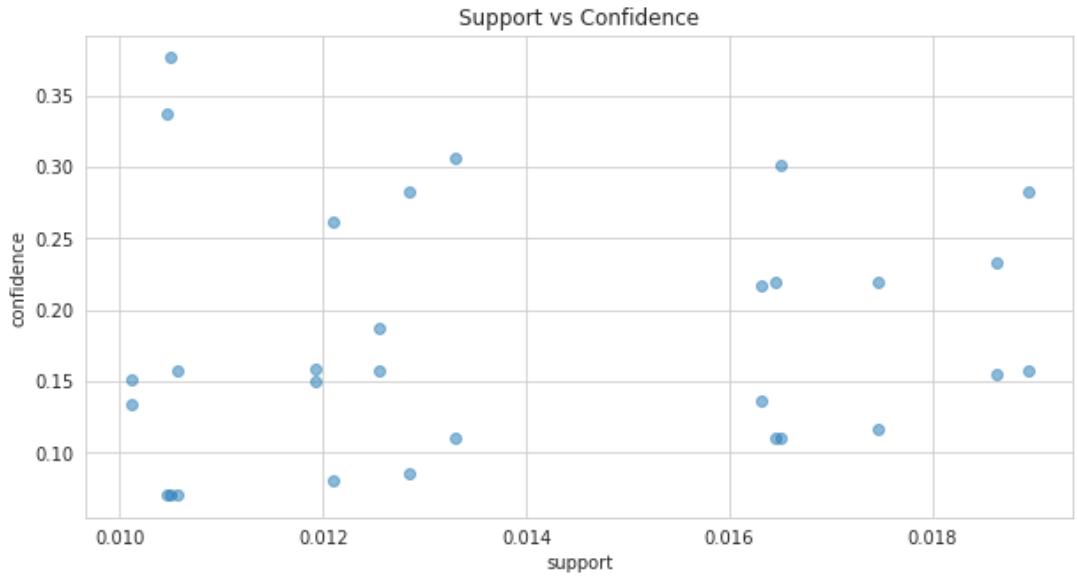

Figure 15 : Support & Confidence values

6.3.Lift value of Associated Rules

The lift index is used to rate the most important rules to calculate the (symmetric) association between the precedent and the result of the rules extracted. The lift of an association rule lift (A -> B) = c(A -> B)}/s(B) = s(A -> B)/s(A)s(B), where s(A -> B) and c(A -> B) are the support and confidence of the rule, respectively, and s(A) and s(B) are the supports of the prior and resultant rule. If lift (A , B) = 1 is not associated with itemsets A and B, i.e., they are statistically independent. Lifting values below 1 indicate a negative correlation between itemsets A and B while values above 1 display a positive correlation. Rules having a lift value close to 1 can be of marginal interest.

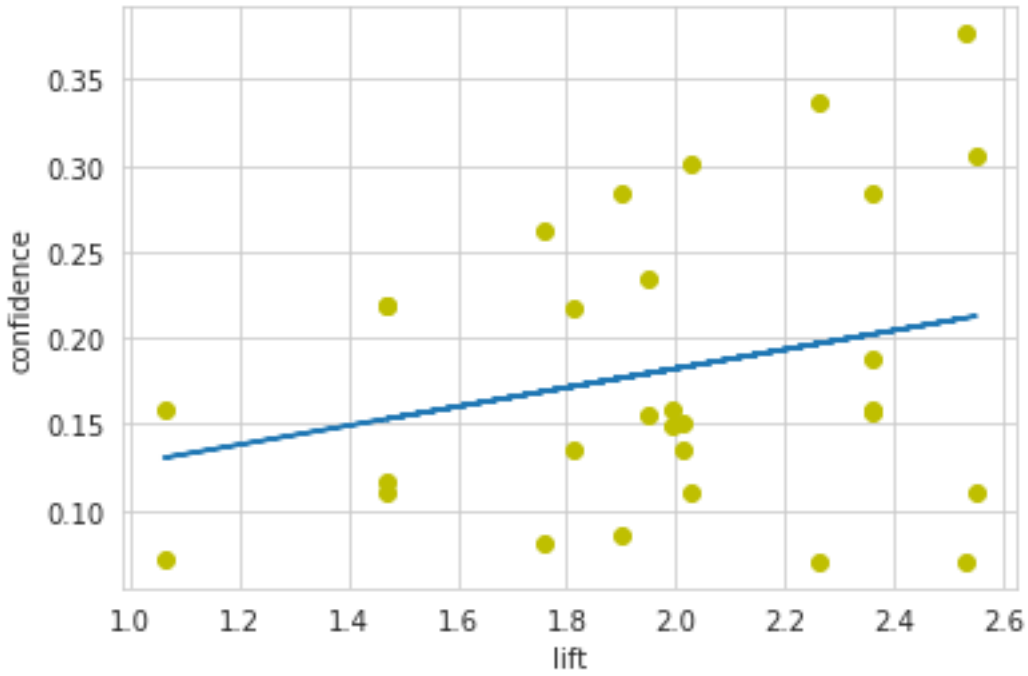

Figure 16 : Confidence & Lift Values

6.4.Network graph

NetworkX (a Python program for the development and analysis of complex networks) to construct a network graph to test the relation between antecedents and consequence obtained after the rule of relation. As can be seen from it, in Instacart, banana is a very popular item which goes well with strawberries, apple, cucumber, apple, avocado and lemon. So, if a person buys one of the last six items, the chances of buying a banana are high (the reverse is also true, i.e. if a consumer purchases a Banana, then the probability of buying one of the 6 items is high).

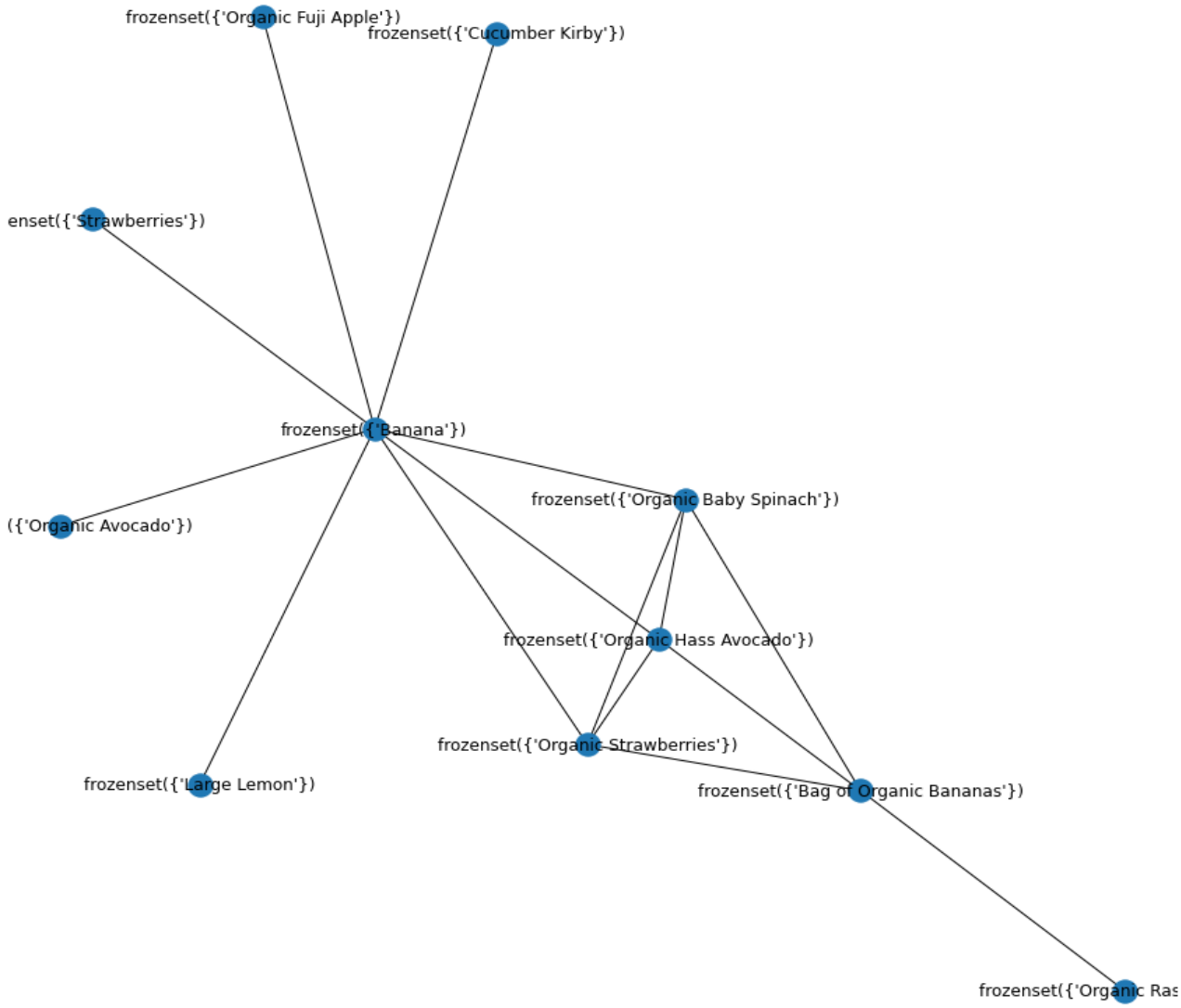

Figure 17 : Network graph of Associated Rules

7. ANALYSIS

7.1.Analysis on Apriori algorithm

Now we are presenting the Apriori algorithm for performance analysis in a comparison process first. For mining items sets, Apriori is the most popular and effective algorithm. The core principle of Apriori is to create multiple passes through data sets or databases that store transactions or data. Apriori algorithm relies on Apriori property which says: "There must be a frequent collection of non-empty item sets. The report also identified a property explaining that all of the super sets fail to pass the test if the system cannot pass the minimum support test. The Breadth First Search (BFS) is used for Apriori algorithm. It also uses downstream locking property (any super collection of an unusually broken item is unusual). The transaction data base is usually horizontally laid out. The frequency of the item set is measured in every transaction.

7.2.Analysis on FP-Growth algorithm

This algorithm uses fragmented and conquering techniques as far as FP growth is concerned, and the structure of FP data is used to achieve a simplified transactional database representation. No regular item sets are required for candidates. The fp tree is mined instead of regular patterns. A list is generated and organized in the decreasing order of support in the initial phase of fp development. A structure named node represents this list. Apart from the root node, each fp node will include the item name, the support count and a pointer connecting to a tree node with a similar item name. These nodes are used for growth of the fp tree. In the creation of the FP tree, common prefixes can be exchanged. The root-to-leaf-node pathways are ordered in a constant order. At this point, when the fp tree is built, frequent patterns are extracted from the FP tree starting in the leaf nodes. Because of the expected layout, FP Growth takes less memory and has effective storage. It's got two crucial moves.

· Contrast a compact fp tree data structure

· Discover frequent fp tree items.

7.3.Comparison between Apriori Algorithm & FP-Growth Algorithm

Evaluation Criteria	Apriori	FP Growth
Generation of Patterns	Apriori creates pattern by combining the items into individual tons, sets, and triplets.	FP growth creates structure by building an FP tree.
Generation of Candidates	Apriori uses the list of candidates.	There is no generation of candidates.
Time	More Time to Execute	Less time compared with algorithm Apriori
Memory Usage	The variations of candidates are kept in memory	A Compact Database Copy is saved
Techniques	Breadth first search	Divide and Conquer

Table 2: Comparison between Apriori Algorithm & FP-Growth Algorithm

8. CONCLUSION

Market Basket Analysis is a conceptual framework that originates in the marketing field and has been used effectively in fields such as bioinformatics, nuclear science, immunology, and geophysics more recently. One reason for MBA's increasing adoption across scientific fields is that by using an inductive approach to theorizing, researchers are able to evaluate the existence of association rules. Considering all, by summing up the whole, we agree that recommendation system can have an efficient impact on marketing and sales research that can be used to make strategic business decisions.

9. FUTURE WORK

Project can be improved by implementing new and advanced mining algorithms along with apriori, fp growth for better performance and fast results for sparse dataset .In the current approach, we only use association rules to exploit the collective information i.e. building an model by finding similarity between customers' products associations and recommending an similar associated item to another customer to purchase. In future work, association rules can also be used to exploit the content-based information i.e. finding a similarity between products, and recommending an products based on interest of a similar products. Content based recommendation system is not based on a lot of user data since the calculation of similarities takes place at the product level. Perhaps we can build recommendation system in future work, incorporating the two approaches into a hybrid approach that can benefit from the strengths of both item-based and customer-based approaches. This application can be extended to other areas such as: sales tracking, product tracking, discount and calculation of prices etc. This method can be applied in future to very large databases where memory space is valuable and needs enhancement. It can be further tuned for improved efficiency and performance.

10. BIBLIOGRAPHY

Agrawal, R., and R. Srikant. 1994. Fast algorithms for mining association rules. In Proceedings of 20th International Conference on Very Large Data Bases, VLDB, Vol. 1215, 487–99. Santiago, Chile: IBM Almaden Research Centre

A Survey of Patients With Self-Reported Severe Food Allergies in Japan by T Imamura 1, Y Kanagawa, M Ebisawa

S. Rangaswamy, Shobha G., "Optimized Association Rule Mining Using Genetic Algorithm," Journal of Computer Science Engineering and information Technology Research (JCSEITR), Vol.2, Issue 1, pp 1-9, 2012.

S. Jain, S. Kabra. "Mining & Optimization of Association Rules Using Effective Algorithm," International journal of Emerging Technology and Advanced Engineering (IJETAE), Vol.2, Issue 4, 2012.

Berry and Linoff. Data Mining Techniques for Marketing, Sales and Customer Relationship Management (second edition), Hungry Minds Inc., 2004

Julander. Basket Analysis: A New Way of Analyzing Scanner Data. International Journal of Retail and Distribution Management, V (7), pp 10-18

S. Erpolat, "Comparison of Apriori and FP-Growth Algorithms on Determination of Association Rules in Authorized Automobile Service Centres," Anadolu Univ. J. Soc. Sci., vol. 12, no. 2, pp. 137– 146, 2012.

I. Cil, "Consumption universes based supermarket layout through association rule mining and multidimensional scaling," Expert Syst. Appl., vol. 39, no. 10, pp. 8611–8625, 2012.

Y. L. Chen, K. Tang, R. J. Shen, and Y. H. Hu, "Market basket analysis in a multiple store environment," Decis. Support Syst., vol. 40, no. 2, pp. 339–354, 2005.

Moore, J. (2012, June 21). Market basket analysis: A powerful tool for gaining customer insight.

Madani, S. (2009). mining changes in customer purchasing behaviour.

Nestorov, S., & Jukic, N. (2003). Ad-Hoc Association-Rule Mining within the Data Warehouse. Hawaii: 36th Hawaii International Conference on System Sciences.

Russell, G. J., & Petersen, A. (2000). Analysis of Cross-Category Dependence in Market Basket Selection. Journal of Retailing, 76(3), 367-392.

relations, p. (2011, 4 23). directory of government retailing system. Retrieved from etka: http://ektad.blogfa.com/post-112.aspx

Bell, D. R., & Boztu, Y. (2007). The positive and negative effects of inventory on category purchase: An empirical analysis. Marketing Letters, 18, 1-14.

Andrews, R. L., & Currin, I. S. (2002). Identifying segments with identical choice behaviours across product categories: An Inter Category Logit Mixture model. International Journal of Research in Marketing, 19, 65-79

Gupta, R. (2013). Finding what to Sell Next to your Customer. Retrieved july 25, 2014

Agrawal, R. 'Fast Algorithms for Mining Association Rules'. p. 13.

Aulakh, R.K. (2015) 'Optimized Association Rule Mining with Maximum Constraints Using Genetic Algorithm'. 3, p. 10.

Berry, M.J.A. and Linoff, G. (2004) *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. 2nd ed. Indianapolis, Ind: Wiley Pub.

Chen, Y.-L. *et al.* (2005) 'Market Basket Analysis in a Multiple Store Environment'. *Decis. Support Syst.* DOI: 10.1016/j.dss.2004.04.009.

Cil, I. (2012) 'Consumption Universes Based Supermarket Layout through Association Rule Mining and Multidimensional Scaling'. *Expert Systems with Applications*, 39, pp. 8611–8625. DOI: 10.1016/j.eswa.2012.01.192.

EconPapers: Anadolu University Journal of Social Sciences. Available at: https://econpapers.repec.org/article/andjournl/ (Accessed: 11 June 2020).

Kanagawa, Y. *et al.* (2009) 'Association Analysis of Food Allergens'. *Pediatric Allergy and Immunology : Official Publication of the European Society of Pediatric Allergy and Immunology*, 20, pp. 347–52. DOI: 10.1111/j.1399-3038.2008.00791.x.

Madani, S. (2009) *Mining Changes in Customer Purchasing Behavior : A Data Mining Approach*. Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-52716 (Accessed: 12 June 2020).

Mostafa, M.M. (2015) ‘Knowledge Discovery of Hidden Consumer Purchase Behaviour: A Market Basket Analysis’. *International Journal of Data Analysis Techniques and Strategies*, 7(4), pp. 384–405. DOI: 10.1504/IJDATS.2015.073867.

Singh, A. and Sinwar, D. (2017) ‘Optimization of Association Rule Mining Using FP_Growth Algorithm with GA’. *International Journal for Research in Applied Science and Engineering Technology*, V. DOI: 10.22214/ijraset.2017.4155.