Product Recommendation System using Associative Rule Mining

Dhrumil Makwana
University of Colorado Boulder
Boulder Colorado USA
dhrumil.makwana@colorado.edu

Tejas Kaushik
University of Colorado Boulder
Boulder Colorado USA
tejas.kaushik@colorado.edu

Abhinav Venkatesh
University of Colorado Boulder
Boulder Colorado USA
abhinav.venkatesh@colorado.edu

ABSTRACT

Through this project we intend to identify customer purchasing patterns by analyzing the items purchased in sequence to determine cross sell. For this project we will consider data from multiple platform store to get relevant interesting patterns for that store platform.

By Identifying this interesting pattern, we can increase the sales by optimizing the product placement, offer special deals and creating product bundle to encourage further sales of these combination. Popularly used in Amazon, Walmart, Target, and many more.

In this project, a recommendation system was built using the Apriori algorithm and by applying association rules. As a result, it was possible to determine items frequently bought together by customers. Given an item as the input, our project can provide the most suitable items as recommendation based on the purchasing patterns observed as per the transaction history of different stores. This project can assimilate data from multiple stores and determine the association rules for the items.

KEYWORDS

Decision making, data mining, association rule, market basket analysis

1 INTRODUCTION

The process of obtaining information from vast amounts of data is known as data mining. Even a phrase as simple as "knowledge mining" could not fully capture the significance of sifting through a lot of data. Mining is the process of extracting a few precise nuggets from a large quantity of raw material. Many other terms, including knowledge mining from databases, knowledge extraction, data/pattern analysis, and data archaeology, have meanings that are somewhat like or dissimilar from data mining.

Simply put, knowledge discovery from databases requires data mining as a necessary step. The steps involved in knowledge discovery as a process are listed below in an iterative sequence:

I) Data cleaning (to eliminate ambiguous and erratic data).

II) Data integration is the process of combining data from various sources.

iii) Data Selection (where data from the database that are pertinent to the analysis task are retrieved).

iv) Data transformation (where data are transformed or consolidated into forms suitable for mining, for example, using summarization or aggregation operations.

v) Data mining (a crucial procedure that employs clever techniques to extract data patterns).

vi) Pattern evolution (to find the patterns that best represent knowledge based on some interesting measures).

vii) Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

For discovering associations between a sizable collection of data elements, association rule mining (ARM) is used. Numerous industries are getting concerned about mining association rules from their databases because to the vast amount of data contained in databases. Examples include catalog design, cross-marketing, and other corporate decision- making processes that can benefit from the discovery of intriguing association patterns between vast quantities of business transaction data. Market basket analysis is a classic illustration of association rule mining.

By finding relationships between different things that customers put in their shopping baskets, this approach analyzes the purchasing behaviors of consumers. By learning which products customers regularly buy with

one another, businesses can use the identification of such links to broaden their marketing strategy.

Market basket analysis (MBA) is a data mining method for identifying relationships across datasets. These associations may be represented by association rules. The official formulation of the problem is as follows: Let I be a collection of things (i1, i2, ..., i nth). Let C be a collection of transactions that results in R. An identifier called RID is used to specifically identify each transaction.

The approach can be expressed as follows: if there are subsets of product items, A and B, then an association rule is in the shape of A*B. It indicates that if a consumer buys A, he or she will also buy B. Support and confidence are two metrics that show how certain newly found association rules. Confidence measures the accuracy of rules.

For Example, if a customer buys diapers they will also tend to buy beer at the same time is represented in association rule.
Diaper - Beer Support = 25% and confidence = 90%

If association rules meet both a type of equation minimum support criterion and a minimum confidence threshold that can be specified by users or domain consultants, they are deemed to be useful. Figure 1 displays an example of a market basket analysis. This is the ideal illustration of association rule mining. This market basket analysis system will assist managers in understanding the groups of goods that customers are most likely to buy. The consumer transaction data from all retail shops may be used in this analysis. Their planning for marketing or advertising strategies will be guided by these findings. Managers can recommend novel arrangements for store layouts, for instance, with the aid of market basket analysis. Things that are frequently bought together can be put close together based on this data to further encourage the selling of such items as a set.

2 LITERATURE SURVEY

The application of market basket analysis can be traced back to [1] where the problem of mining association rules was first explored. With a large dataset of customer transactions wherein each transaction contained items purchased by a customer in one visit to the store, they were able to present an algorithm to identify association rules between items in the dataset. The goal was to identify rules such as "90 percent of transactions that purchase bread and butter also purchase milk." They presented the importance of finding rules with minimum support and confidence. This was further expanded upon by
[2] where Apriori algorithm was formally presented. This was done to help provide a fast algorithm to identify relevant rules by taking decisions based on prior information which was already available. The application of market basket analysis can be applied to the real world as showcased by [3]. In this paper, the customer behavior is analyzed to help companies gain a competitive edge by arranging their products in the supermarket in a way that would lead to an increase in their profits. Apart from the application as presented by [3], data mining techniques also have a variety of applications in different domains. [4] presents an excellent analysis of the application of such techniques in the fields of marketing, web analysis and finance. The effectiveness of different algorithms for associative rule mining is compared in
[5] where matrix Apriori and FP growth is compared in terms of performance. This case study compares the performance on datasets having different characteristics and tries to identify the underlying causes for difference in performance. They found matrix Apriori to perform better than FP-Growth in terms of performance for threshold values below 10% but building the matrix imposed a higher cost which would later help by finding the item sets at a faster rate.

3 DATA SET

Market basket analysis (Dataset 1)

URL:
https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis

Market Basket Dataset contains transaction from 2010 to 2011. In this dataset we will be considering Invoice Number, Item Name and Date. Going forward, this dataset would be referred to as Dataset 1.

Online retail UCI (Dataset 2)

URL:
https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci

This dataset 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 giftware. Many customers of the company are wholesalers. The important attributes used for market basket analysis include the invoice number i.e., the unique 6-digit integral number uniquely assigned to each transaction, description i.e., product name and Invoice date and time. Going forward, this dataset would be referred to as Dataset 2.

4 MAIN TECHNIQUES APPLIED

The Data Cleaning

Remove irrelevant Data like Country, Quantity, Customer ID, Price, Stock Code etc. These parameters won't affect the result of the knowledge extracted after data mining. In case of Country, the dataset is skewed towards 1 country and there are about 32k outliers which has been removed to get higher accuracy. Missing data or unspecified data has been removed as well. After removing irrelevant data, the dataset had the following attributes: Invoice Number/Bill Number, Item Name, Date. As in this project we have 2 datasets which will be merged there might be some redundant data which has been handled. Additionally, a subroutine was created to clear space in front and back of the Item names to remove duplication in data set.

Data Preprocessing

It was observed that in Dataset 1 the attributes did not include a unique code for the products. Dataset 2 on the other hand had an attribute named StockCode for this purpose.

Therefore, as part of preprocessing of the Datasets, a new attribute called HashCode was computed. This attribute was computed by applying the md5 hash on the item description of every product. This action was performed on both the datasets. In this way, a common attribute would act as a unique identifier for the products.

In order to compute the md5 hash, the hashlib library was leveraged. Additionally, data preprocessing was also applied to ensure that the attributes of the two datasets would have common names. Initially Dataset 1 had the following attributes: BillNo, Itemname, Quantity, Date. This was renamed to Invoice, Itemname, Quantity, Date. Similarly, Dataset 2 had the following attributes: Invoice, Description, Quantity, InvoiceDate which in turn was renamed to Invoice, Itemname, Quantity, Date.

Data Integration

The goal of data integration was to combine the two datasets into a single dataset. As mentioned in the data preprocessing section, the attributes of the two datasets were modified in such a way that the two datasets now contained common attributes. These 2 datasets were then combined into a single dataset. The result of this step could be leveraged to apply the apriori algorithm. The output of this step resulted in entries having the following attributes: Invoice, Itemname, Quantity, Date, HashCode

Data Transformation

The data transformation step was applied to transform the result of the previous step (Data integration) into an input that is suitable for the apriori algorithm. The input provided to the apriori algorithm contained a matrix of 1s and 0s. The rows of this matrix represented the transactions (Invoice attribute), and the columns of this matrix represented the items (HashCode attribute). This input was calculated by grouping the dataset based on the Invoice attribute followed by the HashCode. While performing the grouping action, the Quantity attribute was used to generate the total quantity of items (items were identified by the HashCode attribute) purchased per transaction (transactions were identified by the Invoice attribute). If this quantity value was greater than 0, then the value in the matrix was set to 1, else it was set to 0.

Design

In this project, Apriori Algorithm and ARM will be used for and working principle is given below.

Finding common patterns, correlations, connections, or causal structures from data sets found in many types of databases, such as relational databases, transactional databases, and other types of data repositories, is the goal of the method known as association rule mining.

The goal of association rule mining, given a set of transactions, is to identify the rules that will allow us to anticipate the occurrence of a particular item based on the occurrences of the other items in the transaction.

Finding the rules that may control relationships and causal objects between sets of things is the goal of association rule mining, a data mining technique.

Therefore, in a particular transaction involving numerous things, it looks for the principles governing how or why such items are frequently purchased in tandem. As an illustration, peanut butter and jelly are frequently purchased together since many people enjoy making PB&J sandwiches.

It turns out that dads are frequently given the responsibility of doing the shopping while the moms are left with the infant, which is why diapers and beer are frequently purchased together.

The principal uses of association rule mining include:

Analyzing the relationship of things purchased in a single basket or single purchase, as in the aforementioned cases, is known as basket data analysis.

Cross-marketing is collaborating with businesses that enhance your own, not with rivals. For instance, it stands to reason that auto dealers and manufacturers engage in cross-promotional activities with the oil and gas industries.

The items in a company's catalog are frequently chosen to work well together so that purchasing one item will prompt a subsequent purchase of another. These goods are therefore frequently complementary or closely connected.

Working Principle

Step 1: Accept the minimum support as minsup and minimum confidence as minconf.

Step 2: Determine the support count for all the item as s.

step 3: Select the frequent items. (Items with s >= minsup)

step 4: The set candidate k-items is generated by 1-extension of the large (k-1) itemsets generated in step3.

step 5: Support for the candidate k-itemsets is generated by a pass over the database.

step 6: Itemset that do not have minsup are discarded and the remaining itemsets are called large k-itemsets

step 7: The process is repeated until no larger item.

step 8: The interesting rules are determined based on the minimum confidence.

Pseudo Code apriori

Join Step: To generate Ck join Lk-1 with itself.

Prune Step: Any (k-1) itemset which is not frequent cannot be a subset of frequent k-1 itemset.

1. L1= {large 1-itemsets};
2. for (k = 2; Lk-1 ≠ Ø; k++) do begin
3. Ck = apriori - gen (Lk-1);
4. For all transactions t ϵ D do begin
5. Ct = subset (Ck, t)
6. for all candidates c ϵ Ct do
7. c.count++
8. end
9. Lk = {c ϵ Ck | c.count ≥ minsup}
10. end
11. Answer = Uk Lk

k-itemset – An itemset with k items

L_k – Set of large itemsets having k items. Every member Of this set has two parts:-
1.Itemset 2. Support count

C_k – Set of candidate itemsets having k items.Every Member of this set has two parts:-
1.Itemset 2.Support count

5 EVALUATION METHODS

Two key factors—support and confidence—are used to assess an association rule's strength. How frequently a specific rule appears in the database being mined is referred to as support. The frequency with which a given rule holds true in practice is referred to as confidence. A rule may exhibit a strong connection in a data set because it occurs frequently, but it may not occur as frequently when put into practice. In this situation, there would be great support but little confidence.

On the other hand, a rule might not seem to stand out much in a data collection, but further investigation reveals that it happens quite frequently. In this situation, there would be enormous confidence but little backing. These metrics enable analysts to distinguish between correlation and causation and to appropriately assess a particular rule.

The ratio of confidence to support is a third value component, also referred to as the lift value. There is a negative correlation between the datapoints if the lift value is negative. A positive correlation exists if the value is positive, and if the ratio is equal to 1, there is no correlation.

6 TOOLS

1. Python
2. Pandas
3. NumPy
4. Mlxtend - Apriori
5. Mlxtend – Association rules

7 KEY RESULTS

Initial Dataset

BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	255	17850	United Kingdom
536365	WHITE METAL LANTERN	6	01-12-2010 08:26	339	17850	United Kingdom
536365	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	275	17850	United Kingdom
536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	339	17850	United Kingdom
536365	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	339	17850	United Kingdom
536365	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010 08:26	765	17850	United Kingdom
536365	GLASS STAR FROSTED T-LIGHT HOLDER	6	01-12-2010 08:26	425	17850	United Kingdom
536366	HAND WARMER UNION JACK	6	01-12-2010 08:28	185	17850	United Kingdom
536366	HAND WARMER RED POLKA DOT	6	01-12-2010 08:28	185	17850	United Kingdom
536367	ASSORTED COLOUR BIRD ORNAMENT	32	01-12-2010 08:34	169	13047	United Kingdom
536367	POPPY'S PLAYHOUSE BEDROOM	6	01-12-2010 08:34	21	13047	United Kingdom
536367	POPPY'S PLAYHOUSE KITCHEN	6	01-12-2010 08:34	21	13047	United Kingdom
536367	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	01-12-2010 08:34	375	13047	United Kingdom
536367	IVORY KNITTED MUG COSY	6	01-12-2010 08:34	165	13047	United Kingdom
536367	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	01-12-2010 08:34	425	13047	United Kingdom
536367	BOX OF VINTAGE JIGSAW BLOCKS	3	01-12-2010 08:34	495	13047	United Kingdom
536367	BOX OF VINTAGE ALPHABET BLOCKS	2	01-12-2010 08:34	995	13047	United Kingdom
536367	HOME BUILDING BLOCK WORD	3	01-12-2010 08:34	595	13047	United Kingdom

Dataset 1

Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer	Country
489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	01-12-2009 07:45	6.95	13085	United Kingdom
489434	79323P	PINK CHERRY LIGHTS	12	01-12-2009 07:45	6.75	13085	United Kingdom
489434	79323W	WHITE CHERRY LIGHTS	12	01-12-2009 07:45	6.75	13085	United Kingdom
489434	22041	RECORD FRAME 7" SINGLE SIZE	48	01-12-2009 07:45	2.1	13085	United Kingdom
489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	01-12-2009 07:45	1.25	13085	United Kingdom
489434	22064	PINK DOUGHNUT TRINKET POT	24	01-12-2009 07:45	1.65	13085	United Kingdom
489434	21871	SAVE THE PLANET MUG	24	01-12-2009 07:45	1.25	13085	United Kingdom
489434	21523	FANCY FONT HOME SWEET HOME DOORMAT	10	01-12-2009 07:45	5.95	13085	United Kingdom
489435	22350	CAT BOWL	12	01-12-2009 07:46	2.55	13085	United Kingdom
489435	22349	DOG BOWL , CHASING BALL DESIGN	12	01-12-2009 07:46	3.75	13085	United Kingdom
489435	22195	HEART MEASURING SPOONS LARGE	24	01-12-2009 07:46	1.65	13085	United Kingdom
489435	22353	LUNCHBOX WITH CUTLERY FAIRY CAKES	12	01-12-2009 07:46	2.55	13085	United Kingdom
489436	48173C	DOOR MAT BLACK FLOCK	10	01-12-2009 09:06	5.95	13078	United Kingdom
489436	21755	LOVE BUILDING BLOCK WORD	18	01-12-2009 09:06	5.45	13078	United Kingdom

Dataset 2

1). Data Cleaning:

We divided data cleaning in 4 parts.

Step 1: Cleaning rows which does not have country as United Kingdom. As the dataset is skewed towards UK (1.4M data entries)

Results of Step 1:

- Before this step the dataset had [522064 rows x 7 columns] (dataset 1) and [1067371 rows x 8 columns] (dataset 2)
- After performing this step, the result has [487622 rows x 7 columns] (dataset 1), [981330 rows x 8 columns] (dataset 2).
- After this step combine 120483 rows were removed.

Step 2: Removing any rows in which relevant parameters have Nan or empty cells.

Results of Step 2:

- Before this step the dataset had [487622 rows x 7 columns] (dataset 1), [981330 rows x 8 columns] (dataset 2)
- After performing this step, the result has [486167 rows x 7 columns] dataset 1), [976948 rows x 8 columns] (dataset 2).
- After this step combine 5837 rows were removed.

Step 3: Removing the irrelevant columns and reducing the dimension of the dataset.

Results of Step 3:

- Before this step the dataset had 7 columns in dataset 1 and 8 columns in dataset 2.
- After this step the dataset has 4 columns in both datasets.

Step 4: Clearing space in front and back of the Item Name to remove duplication of data

Results of Step 4:

- Before this step the dataset had [486167 rows x 7 columns] dataset 1), [976948 rows x 8 columns] (dataset 2)
- After this step the dataset has [486167 rows x 7 columns] dataset 1), [976948 rows x 8 columns] (dataset 2)

BillNo	Itemname	Quantity	Date
536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01.12.2010 08:26
536365	WHITE METAL LANTERN	6	01.12.2010 08:26
536365	CREAM CUPID HEARTS COAT HANGER	8	01.12.2010 08:26
536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01.12.2010 08:26
536365	RED WOOLLY HOTTIE WHITE HEART.	6	01.12.2010 08:26
536365	SET 7 BABUSHKA NESTING BOXES	2	01.12.2010 08:26
536365	GLASS STAR FROSTED T-LIGHT HOLDER	6	01.12.2010 08:26
536366	HAND WARMER UNION JACK	6	01.12.2010 08:28
536366	HAND WARMER RED POLKA DOT	6	01.12.2010 08:28
536367	ASSORTED COLOUR BIRD ORNAMENT	32	01.12.2010 08:34
536367	POPPY'S PLAYHOUSE BEDROOM	6	01.12.2010 08:34
536367	POPPY'S PLAYHOUSE KITCHEN	6	01.12.2010 08:34
536367	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	01.12.2010 08:34
536367	IVORY KNITTED MUG COSY	6	01.12.2010 08:34
536367	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	01.12.2010 08:34

Fig 1.1: Dataset 1 After Cleaning

Invoice	Description	Quantity	InvoiceDate
489434	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	01-12-2009 07:45
489434	PINK CHERRY LIGHTS	12	01-12-2009 07:45
489434	WHITE CHERRY LIGHTS	12	01-12-2009 07:45
489434	RECORD FRAME 7" SINGLE SIZE	48	01-12-2009 07:45
489434	STRAWBERRY CERAMIC TRINKET BOX	24	01-12-2009 07:45
489434	PINK DOUGHNUT TRINKET POT	24	01-12-2009 07:45
489434	SAVE THE PLANET MUG	24	01-12-2009 07:45
489434	FANCY FONT HOME SWEET HOME DOORMAT	10	01-12-2009 07:45
489435	CAT BOWL	12	01-12-2009 07:46
489435	DOG BOWL , CHASING BALL DESIGN	12	01-12-2009 07:46
489435	HEART MEASURING SPOONS LARGE	24	01-12-2009 07:46
489435	LUNCHBOX WITH CUTLERY FAIRY CAKES	12	01-12-2009 07:46
489436	DOOR MAT BLACK FLOCK	10	01-12-2009 09:06
489436	LOVE BUILDING BLOCK WORD	18	01-12-2009 09:06
489436	HOME BUILDING BLOCK WORD	3	01-12-2009 09:06
489436	ASSORTED COLOUR BIRD ORNAMENT	16	01-12-2009 09:06
489436	PEACE WOODEN BLOCK LETTERS	3	01-12-2009 09:06
489436	CHRISTMAS CRAFT WHITE FAIRY	12	01-12-2009 09:06
489436	HEART IVORY TRELLIS LARGE	12	01-12-2009 09:06
489436	HEART FILIGREE DOVE LARGE	12	01-12-2009 09:06
489436	FULL ENGLISH BREAKFAST PLATE	16	01-12-2009 09:06

Fig 1.2: Dataset 2 After Cleaning

2). Data Preprocessing:

Data preprocessing was split into 2 parts

Step 1: Update the header of columns in both dataset to common attributes

Step 2: Add MD5 Hashcode for item name in each row.

The resulting dataset has the following attributes: Invoice, Itemname, Quantity, Date, HashCode.

After this step the dataset has [486167 rows x 5 columns] dataset 1), [976948 rows x 5 columns] (dataset 2)

Invoice	Itemname	Quantity	Date	HashCode
536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01.12.2010 08:26	55847e72942d9171c8e5b818e6914e36
536365	WHITE METAL LANTERN	6	01.12.2010 08:26	10356e8f2d4a3cfd5fb2e196de97d914
536365	CREAM CUPID HEARTS COAT HANGER	8	01.12.2010 08:26	eadf42df15e10fa91f37cb89d5ab5169
536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01.12.2010 08:26	c6861275f5db1026fc3caaa782c084d5
536365	RED WOOLLY HOTTIE WHITE HEART.	6	01.12.2010 08:26	63d60c916043b9fe2ab08d332420ba11
536365	SET 7 BABUSHKA NESTING BOXES	2	01.12.2010 08:26	c1213ca948df1325454b57647b942f80
536365	GLASS STAR FROSTED T-LIGHT HOLDER	6	01.12.2010 08:26	d620edea437cc23883553e3721aff94e
536366	HAND WARMER UNION JACK	6	01.12.2010 08:28	88ea6c11cf9f84e4d7a9cc38d2bfe837
536366	HAND WARMER RED POLKA DOT	6	01.12.2010 08:28	e61b7db237aa615f0abcbfd686802418
536367	ASSORTED COLOUR BIRD ORNAMENT	32	01.12.2010 08:34	4154e42792974c791a153945bade7377
536367	POPPY'S PLAYHOUSE BEDROOM	6	01.12.2010 08:34	6e2338527902f0d939f20ad976a99e5d
536367	POPPY'S PLAYHOUSE KITCHEN	6	01.12.2010 08:34	6d030e9ce4022792ae41baf95050946e
536367	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	01.12.2010 08:34	6cac8354ec4dd491ee76573c5527248d
536367	IVORY KNITTED MUG COSY	6	01.12.2010 08:34	88cea7f859705473 86f2eda285108c34
536367	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	01.12.2010 08:34	b87b5bab0c5f7d414d8b6de4cafdb2d0
536367	BOX OF VINTAGE JIGSAW BLOCKS	3	01.12.2010 08:34	9e38f76ba89f35a409c84095bfb45245
536367	BOX OF VINTAGE ALPHABET BLOCKS	2	01.12.2010 08:34	546084653d6f8947bafd62597fecd0dd
536367	HOME BUILDING BLOCK WORD	3	01.12.2010 08:34	179c78e9e65fcff8e5cfeac52c84b9de
536367	LOVE BUILDING BLOCK WORD	3	01.12.2010 08:34	81c30021ae36053af2cacea7a5cb248f
536367	RECIPE BOX WITH METAL HEART	4	01.12.2010 08:34	657538ad8cc086b4866c39c0d5cc1db5
536367	DOORMAT NEW ENGLAND	4	01.12.2010 08:34	f055a8dd2f1bc290f481c8486a7926ce
536368	JAM MAKING SET WITH JARS	6	01.12.2010 08:34	cd46d40d7ce57d8c2848bcb914fa285e

Fig 2.1: Dataset 1 After Preprocessing

Invoice	Itemname	Quantity	Date	HashCode
489434	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	01-12-2009 07:45	182069a7a4fc397364788e00238e99
489434	PINK CHERRY LIGHTS	12	01-12-2009 07:45	c87df0cfcdc3f3c88e3c68311be3a95
489434	WHITE CHERRY LIGHTS	12	01-12-2009 07:45	b1e101d7a4776cd22c3e8a361411b5
489434	RECORD FRAME 7" SINGLE SIZE	48	01-12-2009 07:45	035b037c61b9d95d57ce1637dde59d
489434	STRAWBERRY CERAMIC TRINKET BOX	24	01-12-2009 07:45	aaea8967f52bffc6a93a6154216365
489434	PINK DOUGHNUT TRINKET POT	24	01-12-2009 07:45	47b93763528dc0727f50f2c6bb8174
489434	SAVE THE PLANET MUG	24	01-12-2009 07:45	04bd8a143089e13d2b7bf7fc979503
489434	FANCY FONT HOME SWEET HOME DOORMAT	10	01-12-2009 07:45	67b68dc8db529676e94e03247046c2
489435	CAT BOWL	12	01-12-2009 07:46	fe179a550d92267d9482107a55b4e9
489435	DOG BOWL , CHASING BALL DESIGN	12	01-12-2009 07:46	6aab7ce4fb69dd4f8522c5c541ecd5
489435	HEART MEASURING SPOONS LARGE	24	01-12-2009 07:46	77634e20bf248169c36bf47aeda767
489435	LUNCHBOX WITH CUTLERY FAIRY CAKES	12	01-12-2009 07:46	ceb0325871c52a391c715d2a5df083
489436	DOOR MAT BLACK FLOCK	10	01-12-2009 09:06	50f449afdc1688cb1c864853a8eb0b
489436	LOVE BUILDING BLOCK WORD	18	01-12-2009 09:06	81c30021ae36053af2cacea7a5cb24
489436	HOME BUILDING BLOCK WORD	3	01-12-2009 09:06	179c78e9e65fcff8e5cfeac52c84b9de
489436	ASSORTED COLOUR BIRD ORNAMENT	16	01-12-2009 09:06	4154e42792974c791a153945bade73
489436	PEACE WOODEN BLOCK LETTERS	3	01-12-2009 09:06	9ab666bec2069d5bc50de2f061ad7a
489436	CHRISTMAS CRAFT WHITE FAIRY	12	01-12-2009 09:06	462084eb4b3f4de396400367466ca7
489436	HEART IVORY TRELLIS LARGE	12	01-12-2009 09:06	30f3938cf1d189813d152350b1030e
489436	HEART FILIGREE DOVE LARGE	12	01-12-2009 09:06	ae6a47f53b876f48a1651dbf66556a

Fig 2.2: Dataset 2 After Preprocessing

3). Data Integration:

In data integration step the two datasets were combined into a single dataset

The result has more than 1445614 x 5 data entries combined after data integration.

Invoice	Itemname	Quantity	Date	HashCode
536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01.12.2010 08:26	55847e72942d9171c8e5b818e6914e3
536365	WHITE METAL LANTERN	6	01.12.2010 08:26	10356e8f2d4a3cfd5fb2e196de97d914
536365	CREAM CUPID HEARTS COAT HANGER	8	01.12.2010 08:26	eadf42df15e10fa91f37cb89d5ab5169
536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01.12.2010 08:26	c6861275f5db1026fc3caaa782c084d5
536365	RED WOOLLY HOTTIE WHITE HEART.	6	01.12.2010 08:26	63d60c916043b9fe2ab08d332420ba11
536365	SET 7 BABUSHKA NESTING BOXES	2	01.12.2010 08:26	c1213ca948df1325454b57647b942f80
536365	GLASS STAR FROSTED T-LIGHT HOLDER	6	01.12.2010 08:26	d620edea437cc23883553e3721aff94e
536366	HAND WARMER UNION JACK	6	01.12.2010 08:28	88ea6c11cf9f84e4d7a9cc38d2bfe837
536366	HAND WARMER RED POLKA DOT	6	01.12.2010 08:28	e61b7db237aa615f0abcbfd686802418
536367	ASSORTED COLOUR BIRD ORNAMENT	32	01.12.2010 08:34	4154e42792974c791a153945bade7377
536367	POPPY'S PLAYHOUSE BEDROOM	6	01.12.2010 08:34	6e2338527902f0d939f20ad976a99e5d

Fig 3.1: Dataset After Integration

4). Data Transformation:

In data transformation we are grouping the dataset by Invoice and HashCode attribute. The grouping action was performed by calculating the sum of the Quantity attribute.

The resulting dataset was a matrix of 1s and 0s where the row represents the Invoice, and the columns represent the HashCode of the items

The result has 55367 x 5631 data entries that would next be provided to the apriori algorithm for processing.

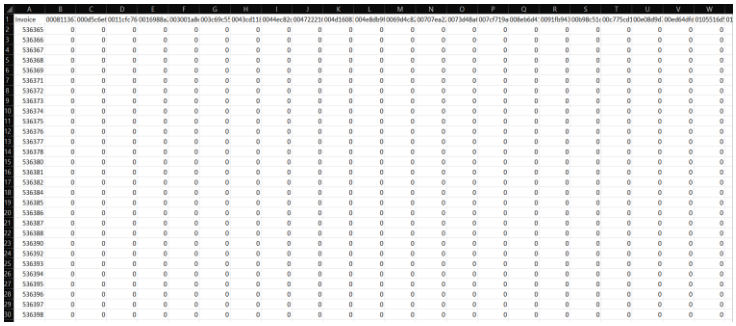

Fig 4.1: Dataset After Transformation

5). Apriori algorithm:

The frequent item sets were computed using the apriori algorithm. The minimum support used was 0.01.

The resulting itemsets were sorted based on the support values in descending order.

The output of the itemsets is shown below

	support	itemsets
205	0.120683	(55847e72942d9171c8e5b818e6914e36)
96	0.082463	(29e30bee90acbea2857f2ef787fe19f5)
383	0.075970	(ac3c46aa4f4822c13368734916814169)
232	0.066393	(63807ffdeb30706f370000821bfc8516)
159	0.062010	(4154e42792974c791a153945bade7377)
127	0.053208	(32eae5f073828207e230b432f0c8f175)
465	0.053027	(d16ac30e35cc2cbf02145487b15eeee7)
99	0.051206	(2a58ffee7f85b680b99e5d96337710a5)
309	0.050755	(8e2a7f29acc72e5365ac27eca251d9af)
56	0.049943	(1851ced19231c9e37b5036ab0d143ab2)
67	0.047761	(1c3ba7b45b8d8f4fc43b0b44f16e644f)
379	0.047454	(aaea8967f52bffc6a93a6154216365b1)

Fig 5.1: Output after applying apriori algorithm

6). Association rules:

The association rules were determined based on the support metric with minimum support of 0.01.

These rules were then sorted based on confidence in descending order.

A consequent and an antecedent (if) make up an association rule (then). An antecedent is something that can be in the data. An object that is discovered along with the antecedent is called a consequent.

By looking for common if-then patterns in the data and utilizing the support and confidence criterion to pinpoint the most significant associations, association rules are generated.

The output of association rules is shown below

	antecedents	consequents	antecedent support	consequent support	support	confidence	1
548	(c9bb4e5985258ac4ef5d96bf5447659b, 3ce0b157826...	(1f23d9ea2c5e063594e69921455eceb7)	0.019281	0.033494	0.017495	0.907390	27.091
542	(29e30bee90acbea28572ef787fe19f5, c9bb4e59852...	(1f23d9ea2c5e063594e69921455eceb7)	0.012626	0.033494	0.011201	0.887143	26.486
546	(1f23d9ea2c5e063594e69921455eceb7, c9bb4e59852...	(3ce0b15782688e4cbdaecc5135596259)	0.020417	0.034179	0.017495	0.856890	25.070
552	(29e30bee90acbea28572ef787fe19f5, c9bb4e59852...	(3ce0b15782688e4cbdaecc5135596259)	0.012626	0.034179	0.010660	0.844286	24.701
89	(c9bb4e5985258ac4ef5d96bf5447659b)	(1f23d9ea2c5e063594e69921455eceb7)	0.024674	0.033494	0.020417	0.827485	24.705
522	(c05f3ec0e4eface25a6fdfc90061ad68, 18f67b2097b...	(a7e5c3a46763d98594a5e281c04ce646)	0.012481	0.029093	0.010155	0.813584	27.964
583	(2fd83514280901cae3cfd0bef6412551, 8e2a7f29acc...	(ac3c46aa4f4822c1336873491 6814169)	0.015024	0.075970	0.012066	0.803121	10.571
536	(29e30bee90acbea28572ef787fe19f5, 3ce0b157826...	(1f23d9ea2c5e063594e69921455eceb7)	0.017089	0.033494	0.013527	0.791139	23.620

Fig 5.1: Output after mining association rules

8 APPLICATIONS

This project can be leveraged for a variety of applications.

The direct application of this project is a recommendation system. Providing suggestions to customers on items that they could be interested in based on the items that they bought. This would help improve sales.

Additionally, these insights help the store in determining the best layout for displaying the items. Items frequently bought together could be placed close to each other to help provide the customer with easy access to these items.

If the stores were to place new orders for items, this knowledge could also be used to see if orders could be placed on items frequently bought together at once.

Providing discounts on prices on combinations of item bought together would further incentivize the customers to buy them and in turn help drive the sales.

9 REFERENCES

[1] R Agrawal, T Imielinski, A Swami, Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, D C 1993 pp 207 216

[2] R Agrawal, R Srikant, Fast algorithms for mining association rules, Proceedings of the 20 th VLDB Conference, Santiago, Chile, 1994 pp 478 499

[3] Raorane AA, Kulkarni RV, Jitkar BD Association Rule Extracting Knowledge Using Market Basket Analysis Research Journal of Recent Sciences 2012 1 2 19 27

[4] I Bose, R K Mahapatra, Business data mining a machine learning perspective, Information and Management 39 2001 211 225

[5] Yıldız B. and Ergenç B., (Turkey) in Comparison of Two Association Rule Mining Algorithms without Candidate Generation, International Journal of Computing, and ICT Research, 674(131), 450-457 (2010)

[6] Chen, D. Sain, S.L., and Guo, K. (2012), Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197-208. doi: [Web Link].

[7] Chen, D., Guo, K. and Ubakanma, G. (2015), Predicting customer profitability over time based on RFM time series, International Journal of Business Forecasting and Marketing Intelligence, Vol. 2, No. 1, pp.1-18. doi: [Web Link].

[8] Chen, D., Guo, K., and Li, Bo (2019), Predicting Customer Profitability Dynamically over Time: An Experimental Comparative Study, 24th Iberoamerican Congress on Pattern Recognition

(CIARP 2019), Havana, Cuba, 28-31 Oct, 2019.

[9] Laha Ale, Ning Zhang, Huici Wu, Dajiang Chen, and Tao Han, Online Proactive Caching in Mobile Edge Computing Using Bidirectional Deep Recurrent Neural Network, IEEE Internet of Things Journal, Vol. 6, Issue 3, pp. 5520-5530, 2019.

[10] Rina Singh, Jeffrey A. Graves, Douglas A. Talbert, William Eberle, Prefix and Suffix Sequential Pattern Mining, Industrial Conference on Data Mining 2018: Advances in Data Mining. Applications and Theoretical Aspects, pp. 309-324. 2018.