

MARKET BASKET ANALYSIS USING DATA MINING TECHNIQUES

V. Vijilesh¹, A. Harini², M. Hari Dharshini³, R. Priyadharshini⁴

¹Assistant Professor, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India

²Student, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India

³Student, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India

⁴Student, Department of Information Technology, Kumaraguru College of Technology, Coimbatore, India

Abstract - The lifeblood of retail businesses has continuously been sales. A distributor will ne'er assume that his customers recognize all of his offerings. But rather, he should create the effort to gift all applicable choices in means will increase client engagement and increase sales. These run on the premise of such innovation having ability to thrill the purchasers with the merchandise, however with such an outsized raft of merchandise leave the purchasers confused of what to buy for and what to not. This is where Machine Learning comes into play, several algorithms are applied for revealing the hidden patterns in data which might be wont to increase sales. Also data processing has application in Retail Industry because it collects great deal of knowledge from customer purchasing history. During this paper we've got used Market Basket Analysis technique by applying association rule mining concepts like Apriori and FP-Growth algorithms which helps the retailers to filter and make recommendations to their customers.

Key Words: Retail sales, Market Basket Analysis, Customer Engagement, Machine Learning, Data Mining.

1. INTRODUCTION

Machine Learning has a good impact on the e-commerce companies that depend on online sales. It's powerful ability to create consistent and more accurate risk assessments, performs predictive tasks and make recommendations for business intelligence purposes. Data processing in retail industry helps to discover the customer buying patterns and trends which contributes to increased quality of customer service, retention and satisfaction. This may help to realize insights into granular behaviour of consumers. Thus we are able to devise strategies to uncover stronger understanding of purchase decisions of the customers. The concept of cross selling will be achieved using these strategies. Cross selling is the ability to sell more number of products to customers by analysing his shopping trends. Cross Selling and Market Basket Analysis techniques are often accustomed analyse and offer additional products to customers as a recommendation with the hope that they would buy benefiting the customer in addition the retail establishment. To realize this, we implemented

association rule mining and generate effective rules that may result in profitable cross selling.

2. METHODOLOGY

There are two major sections:

- Data Pre-processing
- Rules generation by applying Apriori, FP-Growth.

A python program has been developed and implemented in Jupyter Notebook environment. The following packages were imported:

- ❖ Pandas – Loading & Pre-processing of data.
- ❖ Numpy – numeric data calculation.
- ❖ seaborn, matplotlib – Visualization.
- ❖ frequent_patterns – Mining concepts.

2.1 Data Pre-Processing

The dataset is taken from UCI ML Repository which contains transactions of customers in a UK based online retail outfit. The store mainly sells occasional gifts. It contains 8 column fields and 511304 row entries.

This “.csv” is read using Pandas’ “read_csv” function and stored as a data frame.

```

In [1]: import numpy as np
import pandas as pd
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules

In [2]: myretaildata = pd.read_excel('online_retail.xlsx')

In [3]: myretaildata.head()

Out[3]:
InvoiceNo  StockCode  Description  Quantity  InvoiceDate  UnitPrice  CustomerID  Country
0  536365  851234  WHITE HANGING HEART T-LIGHT HOLDER  6  2010-12-01 09:26:03  2356  178510  United Kingdom
1  536365  71803  WHITE METAL LANTERN  6  2010-12-01 09:26:03  2336  178510  United Kingdom
2  536365  844093  CREAM CUPID HEARTS COAT HANGER  8  2010-12-01 09:26:03  2175  178510  United Kingdom
3  536365  842093  KNUITLED UNION FLAG HOT WATER BOTTLE  6  2010-12-01 09:26:03  2336  178510  United Kingdom
4  536365  842092  RED WOOLLY HOTTIE WHITE HEART.  6  2010-12-01 09:26:03  2336  178510  United Kingdom

In [4]: myretaildata['Description'] = myretaildata['Description'].str.strip()

In [5]: myretaildata.dropna(inplace=True)

In [6]: myretaildata['InvoiceNo'] = myretaildata['InvoiceNo'].astype('str')
myretaildata = myretaildata[myretaildata['InvoiceNo'].str.contains('^')]

In [7]: myretaildata.head()

Out[7]:
InvoiceNo  StockCode  Description  Quantity  InvoiceDate  UnitPrice  CustomerID  Country
0  536365  851234  WHITE HANGING HEART T-LIGHT HOLDER  6  2010-12-01 09:26:03  2356  178510  United Kingdom
1  536365  71803  WHITE METAL LANTERN  6  2010-12-01 09:26:03  2336  178510  United Kingdom
2  536365  844093  CREAM CUPID HEARTS COAT HANGER  8  2010-12-01 09:26:03  2175  178510  United Kingdom
    
```

Fig-1:Uploaded Dataset

Before applying algorithms the dataset has to be consolidated by doing the following data preparation steps:

- Dropping all the duplicate entries in every column.
- Strip down extra spaces in “Description” entries.
- Converting Invoice number to String – This is necessary because when we run the analysis, our algorithm will understand invoice number as a string datatype.
- Null values were dropped.

Count of customers for each country is calculated and sorted the count in descending order to see from which country the maximum number of customers purchases. For this analysis, we have filtered the data for one country “Germany”.

```

In [8]: myretaildata['Country'].value_counts()

Out[8]:
United Kingdom    487622
Germany           5042
France            8408
EIRE              2904
Spain            2485
Netherlands      2363
Belgium          2033
Switzerland      1907
Portugal         1501
Australia         1189
Norway            1072
Italy             798
Channel Islands  748
Finland          685
Cyprus            614
Sweden           451
Unspecified      446
Austria          398
Denmark          360
Poland           330
Japan            321
Israel           295
Hong Kong       264
Singapore       222
Ireland          182
USA              179
Canada          151
Greece           145
Malta            112
United Arab Emirates  68
European Community  60
RSA              58
Taiwan           45
    
```

Fig-2: Count of customers for each country

2.1 Market Basket Analysis

Market Basket Analysis(MBA) develops If-Then situation rules, for example, if product X is bought then

item Y is perhaps planning to be bought. The foundations are probabilistic in nature or, similar to that they are derived from (within observations) the frequencies of co-occurrence. Frequency is that the proportion of baskets containing the wishlist. These foundations may be employed in various factors like pricing strategies, product placement, and cross-selling strategies. Association rule mining is employed to come up with rules for Market Basket analysis. This particular algorithm is largely used for recommendation in retail scenario. Any retail scenario, be it online or offline, if we've to try to some cross-sale or up-scale or recommend someone something, this particular method are often very useful. For instance, if someone is buying a chips packet, then from the historical transactions, we all know that individuals who buy chips packet are likely to shop for myonice or sauce etc. So, maybe if someone buys a chips packet, we are able to recommend them to shop for add-ons.

2.3 Association Rule Mining

Association Rules

Bread + Milk + Egg ==> Basket 1

Bread + Milk + Oats ==> Basket 2

Bread + Milk + Wheat ==> Basket 3

"Bread + Milk" as one rule

Association rule learning is a rule-based machine learning technique for locating fascinating relations in massive databases. It's meant to find sturdy rules found in databases .This approach also generates new rules because it analyzes more amount of data.

2.4 Implementation

So all transactions in Germany are taken and grouped by Invoice number and Description and wholly taking the sum of Quantity. This will result in a basket of transactions.

Now we have a basket of data obtained from the data frame. This basket will have “Invoice number” as keys and the name of products as columns. If the value is 0, that product is not present in the corresponding invoice, i.e. transaction. If the value is greater than 0, that product is a part of the corresponding invoice as many times the value.

A function is defined that converts all the non-zero values to 1 in our basket data frame. This is done because the analysis algorithm expects only 0 and 1 as inputs.

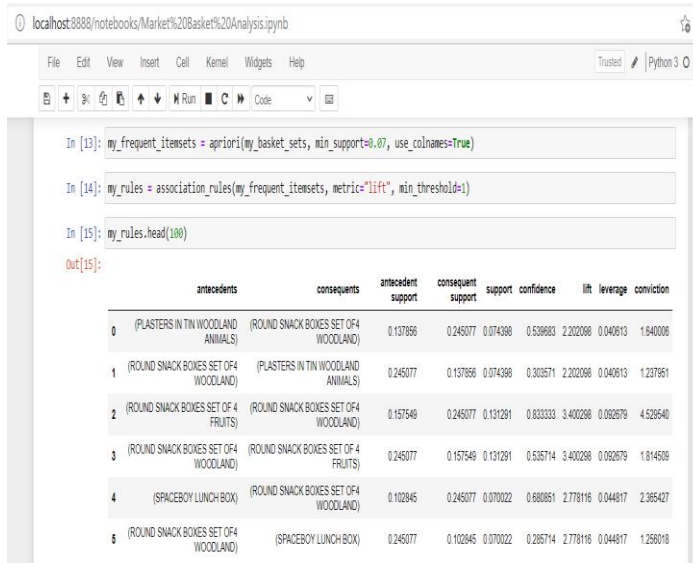
2.5 SUPPORT, CONFIDENCE, LIFT

These specifications are accustomed define the goodness of the foundations generated.

- **Support** is the measure of frequent gathering of things appear together as a percentage of total transactions.
- **Confidence** of the rule is that the ratio of the amount of transactions which include total items additionally to the quantity of transactions that include total items in.
- **Lift** is that the quantitative relation of confidence to expected confidence value.

2.6 Training Model

Next step is to generate frequent itemsets. "apriori" is imported from "frequent_patterns" package. This function is applied on the basket data frame and minimum support value is set. The basket data frame can be now used to generate rules by applying "association_rules" and giving the frequent itemsets as input, choosing lift metric and setting the minimum threshold. The following rules were generated:



```

In [13]: my_frequent_itemsets = apriori(my_basket_sets, min_support=0.07, use_colnames=True)
In [14]: my_rules = association_rules(my_frequent_itemsets, metric="lift", min_threshold=1)
In [15]: my_rules.head(100)
Out[15]:

```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(PLASTERS IN TIN WOODLAND ANIMALS)	(ROUND SNACK BOXES SET OF4 WOODLAND)	0.137669	0.245077	0.074398	0.539883	2.202098	0.040913	1.640006
1	(ROUND SNACK BOXES SET OF4 WOODLAND)	(PLASTERS IN TIN WOODLAND ANIMALS)	0.245077	0.137669	0.074398	0.303571	2.202098	0.040913	1.237891
2	(ROUND SNACK BOXES SET OF4 FRUITS)	(ROUND SNACK BOXES SET OF4 WOODLAND)	0.157549	0.245077	0.131291	0.833333	3.400298	0.082679	4.529540
3	(ROUND SNACK BOXES SET OF4 WOODLAND)	(ROUND SNACK BOXES SET OF4 FRUITS)	0.245077	0.157549	0.131291	0.535714	3.400298	0.082679	1.814509
4	(SPACEBOY LUNCH BOX)	(ROUND SNACK BOXES SET OF4 WOODLAND)	0.102845	0.245077	0.070022	0.680851	2.778116	0.044917	2.365427
5	(ROUND SNACK BOXES SET OF4 WOODLAND)	(SPACEBOY LUNCH BOX)	0.245077	0.102845	0.070022	0.285714	2.778116	0.044917	1.259018

Fig-4: Generation of Frequent item sets

This output says that Item-A implies Item-B with a particular values of support, lift and confidence. These rules can be used to make recommendations. If a combination of products in a basket has a good support, lift and confidence values, those can be recommended to the customers in addition. We can also filter rules based on some conditional scenario as below:

2.7 FP-Growth

FP - Frequent Pattern. It is superior to Apriori algorithm because it does not have to generate all candidate itemsets. It uses the divide-and-conquer strategy and FP-

tree, to seek out frequent itemsets without the need to generate all the itemsets.

COMPARISON APRIORI VS FP-GROWTH

➔ FP-Growth is faster and the runtime increases linearly with increase in itemsets whereas Apriori is slower and the runtime increases exponentially.

➔ FP-Growth scans the database only twice but Apriori scans the database over and over again.

➔ FP-Growth performs no candidate generation but Apriori does.

CONCLUSION

This paper conferred an implementation of cross-selling victimisation Market Basket analysis technique supported knowledge collected from a business on-line retail outfit. The derived insights are useful in deciding for the business wings. we tend to conferred that FP-Growth algorithmic rule is additional economical than Apriori. However the most important bottleneck in any association rule-mining algorithmic rule is that the generation of frequent itemsets. If the dealing dataset has k distinctive product, then probably we've 2k doable itemsets. The generation of rules could be a straightforward method, however computationally pricy, as it grows exponentially with the rise in the set of things. Overall, we tend to get right balance resulting in an inexpensive range of sturdy rules.

REFERENCES

- [1] M. Hossain, A. H. M. S. Sattar and M. K. Paul. "Market Basket Analysis Using Apriori and FP Growth Algorithm," 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2019.
- [2] L. Yongmei and G. Yong. "Application in Market Basket Research Based on FP-Growth Algorithm," 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, USA, 2009.
- [3] Liu, Y., & Guan, Y. (2008). FP-Growth Algorithm for Application in Research of Market Basket Analysis IEEE International Conference on Computational Cybernetics, 269-27. 2008.
- [4] F.Daniel, Customer Segmentation: classification, clustering, marketing. www.kaggle.com