Identifying Customer Purchase Behavior using Apriori Algorithm

Abhinav Duvvuru   
 Artificial Intelligence and Machine Learning  
Lambton CollegeToronto, Canada  
abhinavduvvuru@gmail.com  
  
Sandeep Reddy Alla  
Artificial Intelligence and Machine LearningLambton CollegeToronto, Canada  
itsmealla05@gmail.com  
  
SivaPriya TatiReddy  
Artificial Intelligence and Machine LearningLambton CollegeToronto, Canada  
sivapriya721@gmail.com   
  
Uma Maheswararao Gutta Artificial Intelligence and Machine LearningLambton CollegeToronto, Canada umamaheshgutta@gmail.com

*Abstract*— In this study, we tried to identify the behavior of customers on whether they like and purchase a product or not based on their previous purchases using Apriori and FP Growth Algorithms by applying association rules with the help of confidence, support, and lift parameters. The customer purchase data is cleaned and one hot encoded to feed the algorithm. This study helps in better understanding the customers and their purchase behavior and can increase sales of the products by target promotions and email marketing, etc.

Keywords—Apriori, FP Growth, association rules, frequent itemset, support, confidence, lift, antecedents, consequents, etc.

# Introduction

Now a days, with the help of both retail and online customers data and data mining algorithms, customer purchase experience can be enhanced. Ecommerce and retail companies can target customers based on the customers previous purchases and hidden buying patterns. Companies can recommend the products by email marketing and target promotions etc. In this project, we use the ecommerce data of customers to predict their purchase behaviour and relation between the items by identifying different transactions.

.

# Dataset description

Ecommerce data is an interesting dataset that has information about customer transactions. The dataset used in this project is the data obtained from Kaggle, which has CustomerID, InvoiceNo, StockCode, Description (of product), Quantity, InvoiceDate, UnitPrice, and Country. The dataset has total 8 attributes. Below are the list of attributes and their type:

1. InvoiceNo (Categorical) – Transaction ID of the Customer

2. StockCode(categorical) – Code of the product

3. Description – Product description

4. Quantity – Total quantity of product purchased

5. InvoiceDate – Purchase date

6. UnitPrice – Unit price of the product

7. CustomerID – Customer identification number

8. Country – Country of the customer

The dataset has total 541909 instances and 8 attributes

## Dataset loading

The “data.csv” file can be loaded using Python Pandas and stored as a dataframe. The below figure 1 depicts the data in the dataset.

Table

Description automatically generated

Fig. 1. Overview of dataset

# Data preprocessing

The InvoiceDate column type is converted to datetime type for further analysis.

## Handling missing values

The isnull().sum() method in pandas shows the total number of missing values for each attribute as shown in below figure2.

Chart

Description automatically generated

Fig. 2. Percentage of missing values for each attribute

From the above figure 2, attributes description and customerId has missing values which needs to be handled. Since “Description” is only used for applying the machine learning algorithm, we can handle those values by dropping from the dataset as it has less than 0.5% of missing values.

## Handling error values

The quantity attribute has negative values which should not be possible in real scenario. These values can be dropped form the dataset.

The InvoiceNo attribute has values that starts with “C”, which means cancelled transactions, these values can be dropped form the data.

# Exploratory data analysis and visualization

We applied different visualization libraries like matplotlib and seaborn to analyze the data by using formats like graphs, plots and pie charts.

## Analyzing relation between attributes

From the below figure 3 between country and customers, majority of customers are from United Kingdom followed by Germany and France.

Chart

Description automatically generated

Fig. 3. Bar graph between customers and country

From figure 4, majority of the transactions are observed in United Kingdom and little in Germany and France and null in other countries like Spain, Portugal, and Australia, etc.

Chart

Description automatically generated

Fig. 4. Bar graph between InvoiceId and Country

Chart, pie chart

Description automatically generated

Fig. 5. Pie chart representing top countries based on transactions

## Visualizing top purchased products

Based on the customers purchase of products and its quantity the below figure 6 shows the top 10 products.

Chart, bar chart

Description automatically generated

Fig. 6. Top ten products purchased by customers

The above graph shows that "white hanging heart t-light holder" is the most purchased product followed by "jumbo bag red retrospot" and "regency cakestand 3 tier".

## Visualizing least purchased products

The below figure 7 shows that “dolphin windmill”, “mint diner clock”, “white frangipani hair clip”, “blue gingham rose cushion cover”, etc. are the least purchased products by customers

Chart, bar chart

Description automatically generated

Fig. 7. Least ten products purchased by customers

# model building

After analyzing the transactions of customers, it is evident that 90% of transactions are observed in United Kingdom. So the association rules and machine learning algorithms are applied on this country. Before applying the model, the data needs to be one-hot encoded with on attribute “Description” and index as “InvoiceNo” with values as “sum of quantity. But we only need whether a product is purchased or not.

Below figure 8 depicts the final one-hot encoded values of the dataset

Table

Description automatically generated

Fig. 8. One-hot encoded of Description attribute

## Applying Apriori algorithm

Association rules for the products are applied using Apriori algorithm. The algorithm has mainly three components support, confidence, and lift. Support is referred as popularity of an item based on all the transactions. Confidence refers to the likelihood that an item B is also bought if an item A is bought. Lift is referred as increase in the ratio of sale of B when A is sold.

After converting the data into the specific format, Apriori function is applied, and minimum support is passed to generate frequent item sets and association rules are set on the generated item sets with metric as confidence and minimum threshold as 60% and sort the results based on lift value.

Antecedent represents the left-hand side of the rule and consequent represents the right-hand side of the rule.

Figure 9 show the association rules of products

Application

Description automatically generated with low confidence

Fig. 9. Association rules of products based on Apriori algorithm

Drawbacks of Apriori Algorithm:

* The Apriori algorithm usually takes longer time for candidate generation.
* It needs to process multiple combinations of data which consumes more execution time
* Redundant rules are generated

These issues can be overcome by using FP Growth algorithm

## Applying FP Growth algorithm

It is an efficient algorithm for producing the frequent itemsets without generation of candidate itemsets. It uses a divide and conquer strategy and only need two scans to find out support count. It can mine the items by using lift, confidence, and support by specifying minimum threshold value.

Figure 10 show the association rules of products applied using FP-Growth algorithm.

Text

Description automatically generated with medium confidence

Fig. 10. Association rules of products based on FP-Growth algorithm

## Visualizing support, confidence and lift of algorithms

The below figures 10 and 11 depicts the relation between the confidence and support in terms of lift value. In both the plots confidence is linearly proportional to the lift and when the support value increases the lift value decreases.

Chart, scatter chart

Description automatically generated

Fig. 10. Scatter plot between support and confidence by lift for Apriori algorithm

Chart, scatter chart

Description automatically generated

Fig. 11. Scatter plot between support and confidence by lift for FP-Growth algorithm

# Conclusion

The FP Growth algorithm performs better when compared to Apriori algorithm in terms of execution time of rules and memory utilization.

From the Apriori algorithm with min\_support=0.03, the “Pink regency teacup and saucer” and “Green regency teacup and saucer” itemset has confidence of 0.82 and lift as 15.8. Which shows that out of all transactions that contain pink regency teacup and saucer, 82% of the transactions are likely to contain green regency teacup and saucer. And Green regency teacup and saucer is 15.8 times more likely to be bought by the customers that bought pink regency teacup and saucer, compared to its default sale.

From the FP Growth algorithm with min\_support=0.003, the “Red retrospot peg bag” and “Red retrospot charlotte bag” has confidence of 0.90 and lift as 270. Which shows that out of all transactions that contain red retrospot peg bag, 90% of the transactions are likely to contain red retrospot charlotte bag as well. And the red retrospot charlotte bag is 270 times more likely to be bought by the customers that bought red retrospot peg bag.

##### References

1. “E-commerce dataset”. Kaggle [online dataset], November 2021. Available: Kaggle, <https://www.kaggle.com/carrie1/ecommerce-data>

[Accessed: December 12, 2021]

1. “Complete guide to Association Rules”, towardsdatascience [online document], September 3, 2018. Available: towardsdatascience, <https://towardsdatascience.com/association-rules-2-aa9a77241654>

[Accessed: December 12, 2021]

1. “Association Rule Mining via Apriori Algorithm in Python”. Stackabuse [online document], December 2021. Available: Stackabuse, <https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/> [Accessed: December 12, 2021]
2. “Understanding FP-Growth Algorithm “, [online document], Jul 21, 2020. Available: greatlearning, <https://www.mygreatlearning.com/blog/understanding-fp-growth-algorithm/#:~:text=Apriori%20algorithm%20generates%20all%20itemsets,support%20defined%20by%20the%20user> [Accessed: December 12, 2021]