Ecommerce Data Analysis and Purchase pattern discovery using Apriori Algorithm

By Dayo Samuel, 2022

Abstract

In this project, I analysed ecommerce data to discover patterns in purchase behaviour and product associations in different countries. The goal was to identify potential products that could increase monthly revenue. I used the Apriori algorithm to examine how users select items when making a purchase or adding multiple items to their cart. Through this analysis, I identified patterns in the combination of products that are purchased together and identified opportunities to increase revenue by making targeted product recommendations

Keywords: Ecommerce data analysis , Purchase pattern, Apriori algorithm

2.0 Introduction

Understanding user behaviour and engagement with websites is critical for businesses and organizations to improve the user experience and increase conversions. To effectively analyse user engagement, it is important to consider both the actions users take on the website (such as clicking on links, scrolling, or filling out forms) and the psychological and emotional responses of the users (such as motivation, satisfaction, or frustration). By studying both of these aspects, it is possible to gain a more complete understanding of user behaviour and identify ways to optimize the website to meet the needs and preferences of the users. Additionally, analyzing patterns of association in how users purchase products can provide valuable insights into the decision-making process and help businesses better understand their customers.

According to Łapczyński et al. (2013), modern forms of commerce and sales support systems offer businesses significant potential for analysing purchases. In their study, the authors experiment with building a model of user behaviour on a website and implementing it in a real-time recommendation system. Through this process, they were able to gain insights into consumer buying behaviour in Poland and other European countries.

In a research study conducted by Saucha in 2021, the goal was to find out which customer groups were most likely to respond to certain marketing strategies and how to use Apriori and FP algorithms to gain a competitive advantage. The results showed that these algorithms were effective in determining which products customers were most interested in. This information can be used by businesses to tailor their product offerings and marketing efforts to better meet the needs and preferences of their target customers

Guo, Y et al. (2017), conducted a study with the objective of making mobile e-commerce shopping more convenient and reducing information overload through the use of a recommendation system based on an improved Apriori algorithm. The results of this study showed that this recommendation system, which combines real-time accuracy and recommendation accuracy through data mining, effectively improves the efficiency of mobile e-commerce.

I aim to identify potential products that could boost monthly revenue. To do this, I employ the Apriori algorithm to analyse how users choose items when making a purchase or adding multiple items to their cart. By studying these patterns, I am able to identify combinations of products that are often purchased together and identify opportunities to increase revenue by making targeted product recommendations to users.

2.1 Datasets

The dataset used for this ecommerce data analysis and experimenting purchase pattern of user is gotten from the UCI (UCI Machine Learning Repository: clickstream data for online shopping Data Set), which was credited to Mariusz et al. (2013) as the source. The dataset used in this study consists of information on clickstream data from an online store that sells clothing for pregnant women. The dataset includes 14 variables. The 'Year' variable only includes information from 2008, while the 'Month' variable ranges from April to August. The 'Day' variable represents the day of the month, and the 'Order' variable reflects the sequence of clicks during a single session. The 'Country' variable indicates the origin of the IP address, with categories ranging from Australia to the United Arab Emirates. The 'Session ID' variable includes session identification information, and the 'Page 1 (Main Category)' variable indicates the main product category, with options such as trousers and skirts. The 'Page 2 (Clothing Model)' variable includes information about the code for each product, and the 'Colour' variable represents the colour of the products. The 'Location' variable indicates the location of the product's photo on the page, which has been divided into six sections: top left, top in the middle, top right, bottom left, bottom in the middle, and bottom right. The 'Model Photography' variable has two categories: en face and profile. The 'Price' variable represents the price in US dollars, while the 'Price 2' variable indicates whether the price of a particular product is higher than the average price for the entire product category, with values of 1 for yes and 2 for no. The 'Page' variable represents the page number within the e-store website, which ranges from 1 to 5.

2.3 Explanation and Preparation of dataset

This dataset includes information about different time scales, such as monthly, daily, and yearly. It also includes information about specific categories and subcategories of products, which can be used to focus on specific niche subcategories based on the data points.

The important question, covered as part of this analysis include:

- Which of the products under the Clothing page has the highest purchase?
- Which of the countries has the highest purchase in total?
- Applying APRIORI for the first 6 to 10 product with the highest purchase to discover patterns of user's behavior in the e-shop
- Compare customer buying behavior in United Kingdom and other EU (European Union) countries

Dataset Preparation and Processing:

The dataset used as part of the analysis of ecommerce domain. The snapshot of the data is referenced in Appendix 2. The dataset also has both dependent and independent variables, this identification is done using correlation matrix of the data. And using heatmap from Seaborn to visualise the result as shown in Figure 1

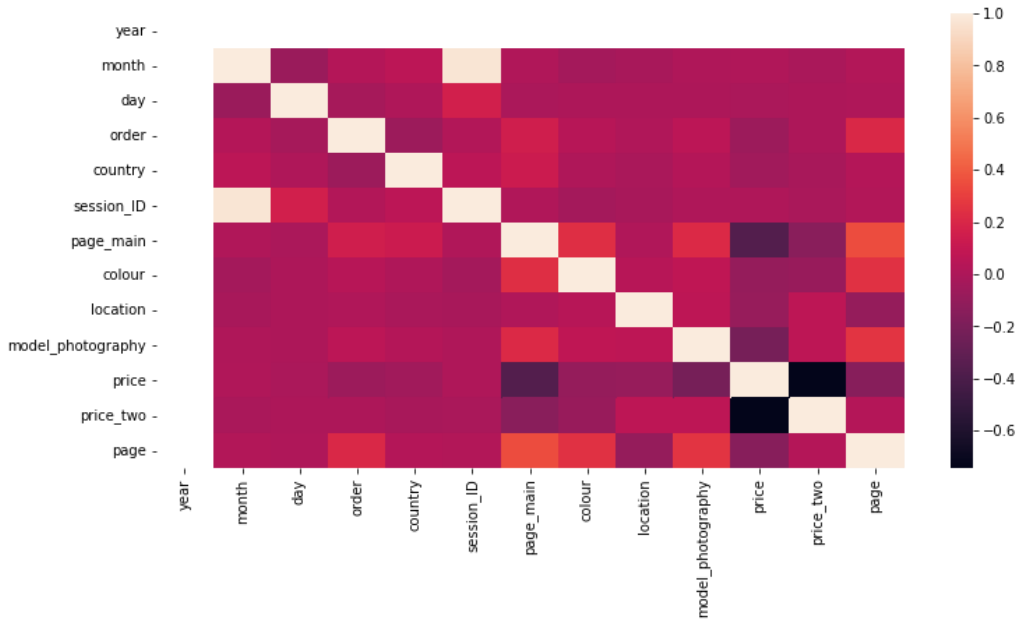

Figure 1 Correlation between features

Independent variables: Year, Month, Day, Session_ID, Page.

Dependent variables: (Country, Location), (Page main, clothing_model_page), (clothing_model_page, colour), (order, price).

Using the function data.isnull().sum() to check null values in my data and it was seen that there were no null values found.

Discovering Unique Pattern and Insight

Using python implementation, 217 unique clothing model were seen on page 2. And monthly, the purchase distribution was checked. And a deeper insight into the purchasing pattern at a subcategory level which takes the clothing model and total amount gained into account was done. Likewise, the country with the highest purchase

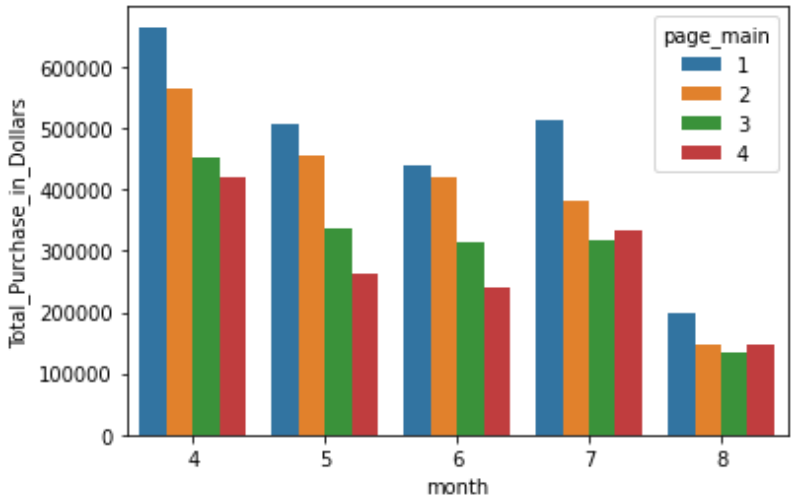

Figure 2 Purchase Distribution pattern at Main category level on a monthly granularity

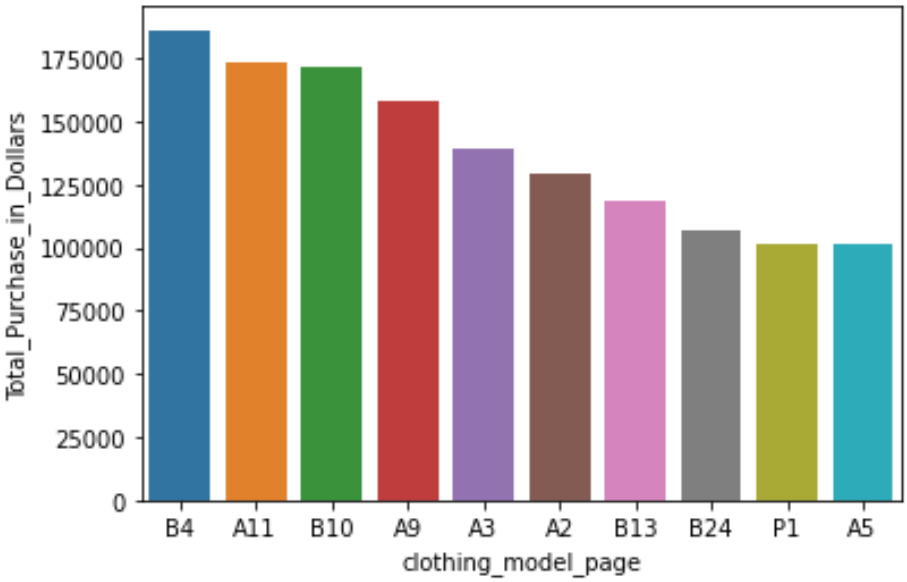

Figure 3 Products having Highest Purchase Overall

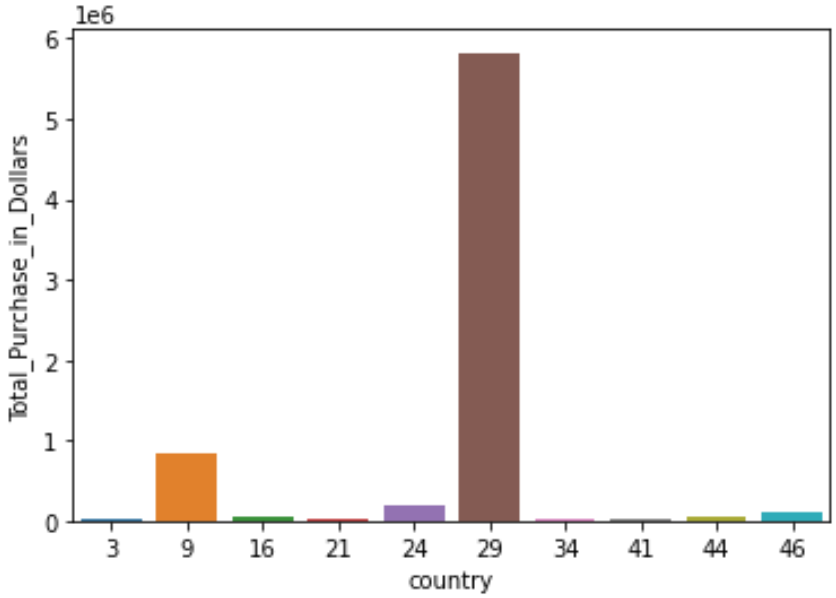

Figure 4. Countries with Highest purchase

2.4 Association Rule Mining

Exploring association rules is a part of data mining and unsupervised learning. And it is the process of finding the rules that may govern associations between set of items. The term market basket analysis refers to monitoring shopping or purchase patterns to increase consumer satisfaction and so increase completed purchases. There are several algorithms that can be used for association rule mining, including the Apriori algorithm and the FP-growth algorithm. These algorithms use different approaches to identify and generate association rules, but they all aim to uncover hidden patterns and relationships in the data.

In this study, and in all the transaction made with different items the analysis tries to find the rules that govern how or why each item are often bought together. In addition, association rules are commonly used for recommender systems and click stream analysis. Having imported all the required libraries including Apriori. This helps to find the most frequent combinations in the database and finds the association rules between the items based on the following 3 factors:

- **Support**: the probability that X and Y come together
- **Confidence**: the conditional probability of Y knowing X. In other words, how often does Y happen when X happened first.
- **Lift**: the ratio between support and confidence. A lift of 2 means that the likelihood of buying X and Y together is 2 times more than the likelihood of just buying Y.

2.5 Experimentation using Association Rule Mining

In practice, a lift of at least 1 is necessary for a rule to be considered relevant, the steps of the Association rule mining done using mlxtend library on the dataset are shown in Appendix 2:

Discovering Pattern of user's behaviour at country level

Prior to the insight found, Poland has the highest purchase history. Finding on the sub-category level to check with products attribute they are more likely to buy. Using Poland as a prototype analysis, I extracted top 10 products with the highest purchase within Poland and extracted records about the same products.

Creating a basket to find pattern and correlation between various products

To identify patterns in product combinations, I created baskets with the following groups: page main and clothing, and the number of orders. This process was repeated for other countries. The Apriori algorithm operates on a set of transactions, where each transaction is a set of items. In order to use this algorithm, the items in each transaction must be represented as numerical data.

In this case, I used one-hot encoding to convert the baskets to a binary format. One-hot encoding is a method used to represent categorical variables in a dataset as numerical data. It involves creating a new binary (0 or 1) column for each unique category in a categorical variable. Values above 1 are converted to 1, and values at 0 or below are converted to 0. This was done primarily for the application of the Apriori algorithm. By using one-hot encoding, I was able to apply the Apriori algorithm to the categorical data and identify patterns and relationships between different categories.

Applying Apriori algorithm to capture user's behaviour when they buy a product:

Here, I build the model with the Apriori function and collect the inferred rules that meet the minimum support of 0.01 using association_rules function.

Similarly, applying Apriori algorithm to use purchase pattern in different countries and experimenting by creating a simple transaction for UK shoppers

2.6 Results analysis and discussion

Based on the analysis, it appears that the purchase patterns in the UK are similar to those in Poland, despite the fact that Poland has a higher number of completed transactions. In particular, in Poland, customers who purchase product A11 are more likely to also purchase products A2 and A3. Similarly, in the UK, shoppers who purchase product A11 are more likely to also purchase products A2 and B4, as well as B10. Table 4 below provides a comparison of customer buying behaviour in the UK and other EU countries. In the Czech Republic and Germany, consumer purchase patterns are largely similar, but the outcomes of those patterns differ. Despite having similar antecedents, or preceding factors, the consequents, or resulting effects, of these patterns in the Czech Republic and Germany are distinct. On the other hand, in Lithuania, the consequent of purchase pattern A11 is present, which also happens to be the antecedent for both the Czech Republic and Germany.

Table 4. Purchase Pattern amongst highest purchase countries

Country	Antecedents	Consequents
Poland	A11	A2
United Kingdom	A11	A2
Czech Republic	A11	A10
Lithuania	A12	A11
Germany	A11	A3

Conclusions

Data mining and unsupervised learning are useful and vital experimental approaches for ecommerce, as they allow for gaining insight into user behaviour patterns on the website and transactional patterns of shoppers using association rule mining. This study and analysis were conducted to provide insight and identify purchase patterns for different customers and potential products. The results of this study showed that the purchase patterns of customers in the United Kingdom are similar to those in Poland. Therefore, similar product recommendations may be effective for customers in both the United Kingdom and Poland, but it is important to note that these findings may not necessarily apply to customers in other countries.

References

Guo, Y., Wang, M. and Li, X. (2017), "Application of an improved Apriori algorithm in a mobile e-commerce recommendation system", *Industrial Management & Data Systems*, Vol. 117 No. 2, pp. 287-303. https://doi.org/10.1108/IMDS-03-2016-0094

Łapczyński, M., & Białowąs, S. (2013). Discovering Patterns of Users' Behaviour in an E-shop-Comparison of Consumer Buying Behaviours in Poland and Other European Countries. *Studia Ekonomiczne*, *151*, 144-153.

S. Diwandari and U. Zaky, "Analysis of Customer Purchase Behaviour using Association Rules in e-Shop," 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2021, pp. 144-149, doi: 10.1109/ICITISEE53823.2021.9655892.

Appendices

Appendix One

(These screenshots depict the referenced procedures that were completed for Task 2. The source of the information is a combination of self-written notes and a Jupyter notebook.)

- Discovering Pattern of user's behaviour at country level

```
country_model_data = pd.DataFrame(data.groupby(['country','clothing_model_page']).\
                        agg(Total_Purchase=('price','sum'))).reset_index().\
                        sort_values(by="Total_Purchase",ascending=False)
country_model_data.head(10)
```

	country	clothing_model_page	Total_Purchase
1805	29	B4	152880
1779	29	B10	144921
1737	29	A11	132308
1777	29	A9	110618
1782	29	B13	100750
1757	29	A3	98712
1746	29	A2	97051
1870	29	P1	95798
1793	29	B24	89148
1768	29	A4	75886

- Extraction of top 10 products with the highest purchase within Poland and extracting records about the same.

```
# Extracting top 10  Purchase Products list
products = list(country_model_data.head(10)['clothing_model_page'].values)
products
```

```
['B4', 'B10', 'A11', 'A9', 'B13', 'A3', 'A2', 'P1', 'B24', 'A4']
```

```
#Extracting Data for top 10 Products for Poland
highest_purchase_model = data[(data['country']==29) & (data['clothing_model_page'].isin(products))]
highest_purchase_model.shape
```

```
(19921, 14)
```

- Creating a basket to find pattern and correlation between various products

```
basket_Poland = (highest_purchase_model.groupby(['page_main','clothing_model_page'])['order']
          .sum().unstack().reset_index().fillna(0)
          .set_index('page_main'))
```

```
basket_Poland.shape
```

```
(3, 10)
```

```
basket_Poland
```

clothing_model_page	A11	A2	A3	A4	A9	B10	B13	B24	B4	P1
page_main										
1	11997.0	11725.0	9633.0	10320.0	9750.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	11143.0	8859.0	13026.0	17767.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20650.0

- Converting this basket to binary format using one-hot encoding

```
#Function to Convert the non-zero value to binary via hot encoding method
def hot_encode(x):
    if(x<= 0):
        return 0
    if(x>= 1):
        return 1

# Encoding the datasets
basket_encoded = basket_Poland.applymap(hot_encode)
basket_Poland = basket_encoded
```

```
basket_Poland
```

clothing_model_page	A11	A2	A3	A4	A9	B10	B13	B24	B4	P1
page_main										
1	1	1	1	1	1	0	0	0	0	0
2	0	0	0	0	0	1	1	1	1	0
4	0	0	0	0	0	0	0	0	0	1

- Applying Apriori algorithm to capture user's behaviour when they buy a product:

```
frq_items = apriori(basket_Poland, min_support = 0.05, use_colnames = True)

# Collecting the inferred rules in a dataframe
rules = association_rules(frq_items, metric ="lift", min_threshold = 1)
rules = rules.sort_values(['confidence', 'lift'], ascending =[False, False])
print(rules.head())
```

```
  antecedents consequents  antecedent support  consequent support   support  \
0       (A11)        (A2)            0.333333            0.333333  0.333333
1        (A2)       (A11)            0.333333            0.333333  0.333333
2       (A11)        (A3)            0.333333            0.333333  0.333333
3        (A3)       (A11)            0.333333            0.333333  0.333333
4        (A4)       (A11)            0.333333            0.333333  0.333333

   confidence  lift  leverage  conviction
0         1.0   3.0  0.222222         inf
1         1.0   3.0  0.222222         inf
2         1.0   3.0  0.222222         inf
3         1.0   3.0  0.222222         inf
4         1.0   3.0  0.222222         inf
```

- Applying Apriori algorithm to use purchase pattern in different countries and experimenting by creating a simple transaction for UK shoppers

Experimentation and Creating a simple transaction for UK shopper

```
import utils
%matplotlib inline
```

```
from apyori import apriori
import utils
```

```
#creating simple transactions
transactions =[
    ['A11', 'A2', 'B4', 'B10'],
    ['A8', 'B13', 'B24'],
    [ 'A3', 'B11', 'B31'],
    ['A2', 'B4', 'B10'],
]
transactions
```

```
[['A11', 'A2', 'B4', 'B10'],
 ['A8', 'B13', 'B24'],
 ['A3', 'B11', 'B31'],
 ['A2', 'B4', 'B10']]
```

```
#Generating rules
Rules=list(apriori(transactions,min_support=0.2,min_confidence=0.5))
```