BOOSTING ONLINE SALES USING ASSOCIATION RULE MINING

BY

DABERECHI CORNELIUS AHANONU

(UNIVERSITY OF SALFORD)

Introduction

Association rule mining is a technique that identifies frequent items, correlations, associations, or causal structures in data sets found in a database, including relational databases, transactional databases, and other types of repositories (Nguyen, 2020). The rule shows how frequently an itemset occurs in a transaction. It allows retailers to identify relationships between the items that people buy together frequently (Pei., Han., and Mao, 2000). This will help retailers identify new set of opportunities for products they're selling to their customers. The aim is for a given a set of transactions; we want to derive rules that allow us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction. Therefore, given the online retail dataset, we attempt to identify the rules that govern how or why such items are frequently purchased together.

Datasets

The dataset is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The customers are mainly wholesalers from selected parts of the world.

Explanation and preparation of datasets

The data has the following attributes:

Invoice No: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
Stock Code: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
Description: Product (item) name. This contains the combination of the items bought by the customers.
Quantity: The quantities of each product (item) per transaction. Numeric.
Invoice Date: Invoice Date and time. Numeric, the day and time when each transaction was generated.
Unit Price: Unit price. Numeric, Product price per unit in sterling.
Customer ID Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
Country: Country name. Nominal, the name of the country where each customer resides.

The dataset contains over 500000 records with 38 different countries, therefore, to reduce this, we only took some samples by considering transactions carried out by customers in the

United Kingdom. For each transaction, we converted them to a binary form where '1' means that the item was purchased and '0' means item was not purchased.

Description of the algorithms used

Association Mining

Association mining allows us to identify patterns that exists between items in a database (Rygielski., Wang., and Yen, 2002). With association mining, we can determine the probability of one of more items being bought together by a customer. For example, the association rule, {laptop} => {mouse, keyboard} says that a customer who bought a laptop has the high probability of buying mouse and keyboard also. The transaction in the database is usually large, therefore, an efficient algorithm is needed to discover useful information. The association rule process involves two stages: first is to find the frequent itemset; the second is the use the frequent itemset to generate the association rules. An example of an algorithm that can help generate these rules is known as the apriori algorithm.

Association mining can be represented in the following form:

Let I = $\{i_1, i_2, \ldots, i_n\}$ be a set of 'n' binary attributes called items. Let D= $\{d_1, d_2, \ldots, d_n\}$ be set of transaction called database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of form X->Y. The set of items X and Y are called antecedent and consequent of the rule respectively.

Apriori Algorithm

The apriori algorithm was proposed by Rakesh Agrawal and Ramakrishnan Srikant in 1994. The algorithm finds frequent itemsets in a transactional database and models the association between the items. To effectively construct the association rules, we use the following metrics, support, confidence and lift.

Support

The support supp(X) of an item set X is defined as the proportion of transactions in the dataset which contain the item set (Hipp., Güntzer., and Nakhaeizadeh, 2000). This is given by:

$$\text{Support(B)} = \frac{Transactions\ containing\ B}{Total\ transactions}$$

Confidence

Confidence of a rule is defined as conf ($X \Rightarrow Y$) = supp ($X \cup Y$)/supp(X) (Hipp., Güntzer., and Nakhaeizadeh, 2000). Confidence can be interpreted as an estimate of the probability P(Y|X), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS. This is given by:

$$Confidenc(A \rightarrow B) = \frac{Transactions\ containing\ A\ and\ B}{Transactions\ containing\ A}$$

Lift

The lift is the ratio between the confidence and support. This is given by:

$$lift(A \rightarrow B) = \frac{\text{support(A and B)}}{\text{support(A)} \times \text{support(B)}}$$

The application of data-mining techniques to selected datasets that you choose using Python.

The method of data mining technique applied is called the association mining technique. We want to identify relationships between one or more items in a transactional database. We used the 'mlxtend' python package for the purpose of this work.

Results analysis and discussion

Table 1: Top 10 rules with highest lift

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
65	(ROSES REGENCY TEACUP AND SAUCER , GREEN REGEN...	(PINK REGENCY TEACUP AND SAUCER)	0.029837	0.029923	0.020984	0.703281	23.503392	0.020091	3.269348
68	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER , GREEN REGEN...	0.029923	0.029837	0.020984	0.701280	23.503392	0.020091	3.247735
69	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENC...	0.039755	0.023240	0.020984	0.527837	22.712470	0.020060	2.068694
64	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENC...	(GREEN REGENCY TEACUP AND SAUCER)	0.023240	0.039755	0.020984	0.902930	22.712470	0.020060	9.892337
66	(PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	(ROSES REGENCY TEACUP AND SAUCER)	0.024559	0.040734	0.020984	0.854419	20.975684	0.019984	6.589245
67	(ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	0.040734	0.024559	0.020984	0.515152	20.975684	0.019984	2.011846
10	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.029923	0.039755	0.024559	0.820768	20.645746	0.023370	5.357558
11	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.039755	0.029923	0.024559	0.617773	20.645746	0.023370	2.537962
60	(ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.040734	0.029923	0.023240	0.570533	19.066999	0.022021	2.258794
61	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.029923	0.040734	0.023240	0.776671	19.066999	0.022021	4.295313

Column 7 of Table 1 reveals the top 10 rules with the highest lift. The information shows that 'Green Regency T eacup and Saucer, Roses Regency Teacup and Saucer and Pink Regency Teacup and Saucer occurs more often t han expected

We can also do same for the support and confidence as shown in Table 2 and 3 respectively.

Table 2: Top 10 rules with highest support

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
20	(JUMBO BAG RED RETROSPOT)	(JUMBO BAG PINK POLKADOT)	0.082489	0.049332	0.033413	0.405057	8.210874	0.029343	1.597914
21	(JUMBO BAG PINK POLKADOT)	(JUMBO BAG RED RETROSPOT)	0.049332	0.082489	0.033413	0.677308	8.210874	0.029343	2.843302
12	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.040734	0.039755	0.029837	0.732497	18.425368	0.028218	3.589667
13	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.039755	0.040734	0.029837	0.750535	18.425368	0.028218	3.845299
32	(JUMBO STORAGE BAG SUKI)	(JUMBO BAG RED RETROSPOT)	0.048097	0.082489	0.029710	0.617699	7.488247	0.025742	2.399971
33	(JUMBO BAG RED RETROSPOT)	(JUMBO STORAGE BAG SUKI)	0.082489	0.048097	0.029710	0.360165	7.488247	0.025742	1.487732
30	(JUMBO SHOPPER VINTAGE RED PAISLEY)	(JUMBO BAG RED RETROSPOT)	0.048225	0.082489	0.027965	0.579876	7.029730	0.023987	2.183907
31	(JUMBO BAG RED RETROSPOT)	(JUMBO SHOPPER VINTAGE RED PAISLEY)	0.082489	0.048225	0.027965	0.339009	7.029730	0.023987	1.439922
42	(LUNCH BAG BLACK SKULL.)	(LUNCH BAG RED RETROSPOT)	0.051758	0.059249	0.025836	0.499178	8.425057	0.022770	1.878412
43	(LUNCH BAG RED RETROSPOT)	(LUNCH BAG BLACK SKULL.)	0.059249	0.051758	0.025836	0.436063	8.425057	0.022770	1.681469

Table 3: Top 10 rules with highest confidence

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction

64	(PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY...	(GREEN REGENCY TEACUP AND SAUCER)	0.023240	0.039755	0.020984	0.902930	22.712470	0.020060	9.892337
65	(PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY...	(ROSES REGENCY TEACUP AND SAUCER)	0.024559	0.040734	0.020984	0.854419	20.975684	0.019984	6.589245
10	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.029923	0.039755	0.024559	0.820768	20.645746	0.023370	5.357558
60	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.029923	0.040734	0.023240	0.776671	19.066999	0.022021	4.295313
13	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.039755	0.040734	0.029837	0.750535	18.425368	0.028218	3.845299
12	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.040734	0.039755	0.029837	0.732497	18.425368	0.028218	3.589667
8	(GARDENERS KNEELING PAD CUP OF TEA)	(GARDENERS KNEELING PAD KEEP CALM)	0.031923	0.038223	0.023027	0.721333	18.871944	0.021807	3.451355
3	(CHARLOTTE BAG PINK POLKADOT)	(RED RETROSPOT CHARLOTTE BAG)	0.028688	0.038520	0.020388	0.710682	18.449475	0.019283	3.323268
66	(ROSES REGENCY TEACUP AND SAUCER , GREEN REGEN...	(PINK REGENCY TEACUP AND SAUCER)	0.029837	0.029923	0.020984	0.703281	23.503392	0.020091	3.269348
67	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER , GREEN REGEN...	0.029923	0.029837	0.020984	0.701280	23.503392	0.020091	3.247735

Column 6 of Table 3 shows the rules with the highest confidence.

Conclusion

In this work, we have successfully drawn a rule for discovering the relationships that exists between itemsets. These relationships will help us understand our customers better, thereby making policies and strategies that will help boost sales.

References

Agrawal, R. and Srikant, R., 1994, September. Fast algorithms for mining association rules. *In Proc. 20th int. conf. very large data bases*, VLDB (Vol. 1215, pp. 487-499).

Chi Nguyen., 2020. Simple Market Basket Analysis with Association Rules Mining. Available at: https://towardsdatascience.com. (Accessed: 12 November, 2022).

Hipp, J., Güntzer, U. and Nakhaeizadeh, G., 2000. Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations* newsletter, 2(1), pp.58-64.

Pei, J., Han, J. and Mao, R., 2000, May. CLOSET: An efficient algorithm for mining frequent closed itemsets. *In ACM SIGMOD workshop on research issues in data mining and knowledge discovery* (Vol. 4, No. 2, pp. 21-30).

Rygielski, C., Wang, J.C. and Yen, D.C., 2002. Data mining techniques for customer relationship management. *Technology in society*, 24(4), pp.483-502.