Association Rule Mining For UK Grocery Store

I. Introduction

In this report, an application of rule mining is exercised on a data set of purchase transactions from a United Kingdom grocery store sourced from Kaggle [1], which will be analyzed to uncover association rules. These association rules will imply the likelihood of customers purchasing grocery items together which can be leveraged to determine the future placement of items within the grocery store to incentivize increased purchases. Careful consideration will be given to item groups based on their values for support, confidence, and lift as well as antecedent and consequent.

The future sections of this report will analyze the data sets validity and usefulness, the application of the apriori algorithm to establishing association rules, a discussion of results yielded from analyzing the data set, and conclude with the possible applications of these insights.

II. Data Description

This data set was chosen and formatted to be well fit for association rule mining, specifically for the framework of Orange Software. The data set has twenty rows recording transactions for a grocery store. The set has eleven features used as the columns where each column represents a specific item that was purchased. The entries in this data table show either a '1' which means the item was purchased or is blank which means the item was not purchased. These are binary values and thus our data only represents associations of a single set of items to another set of items and not a set of sets to another set of sets. Table 1 below shows what the attributes contained within the data set.

TABLE I.

Attribute	Type	Example Value	Description
ITEM	Nominal (string)	"Bread"	Name of the purchased item
COUNT	Numeric (binary)	1	1: Item purchased ? or "": Item not purchased

These eleven attributes all played a part in demonstrating the associations that can exist between items. A frequency distribution for all the attributes was created, see Figure 1, to show the popularity of the grocery items that were supplied in the data set. Some of these items appear far more frequently than others. The impact that grocery items with high frequencies have on the apriori algorithm will be discussed in Section III. The frequency distribution in this instance also represents the number of transactions an item appeared in from the total number of transitions since our count is a binary value akin to yes or no.

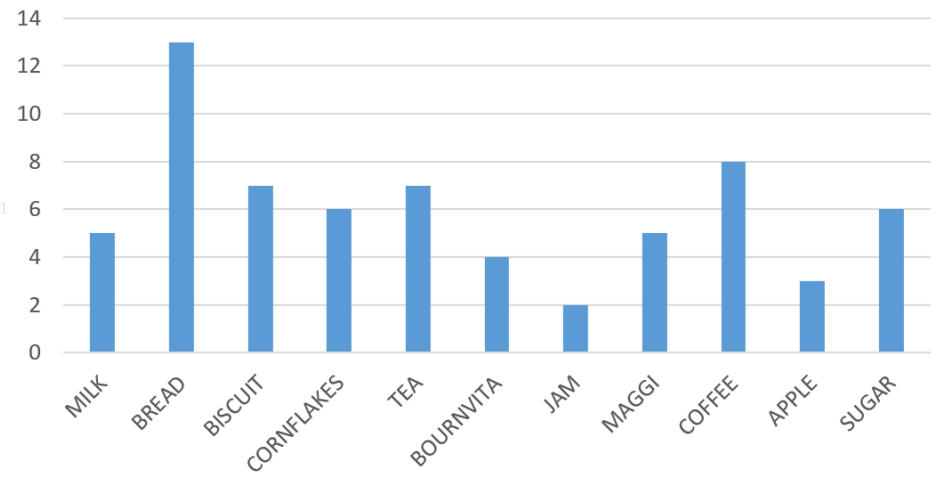


Fig. 1. Frequency distribution of
Purchased Grocery Items

III. METHODOLOGY

Association rule mining was achieved on this data set by applying the apriori algorithm which is a breath-first search based algorithm that determines the support, confidence, and lift of the features associated with certain other features. Support represents how often an item appeared in a transaction and confidence represents the likeliness that one item or set of items will result in the purchase of another item or set of items. The lift is especially significant because simply observing the support and confidence can mislead us to a false conclusion. The lift shows how much more likely a certain item is to be purchased relative to the items general purchase rate. As an example, BREAD appears in over half the transactions recorded thus the feature will have a high support and likely a high confidence, but there is nothing surprising or insightful of bread being a part of an association with another feature. It was therefore determined in this phase to remove BREAD from the transactions. The result is recorded in Figure 2.

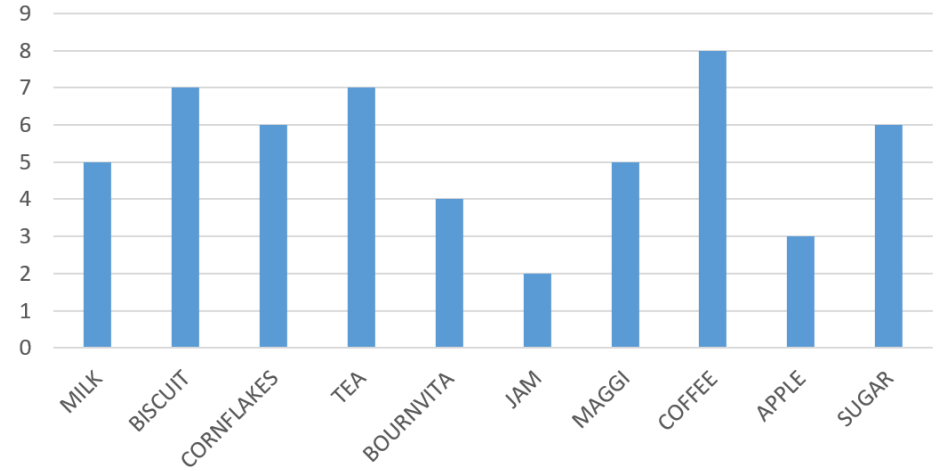


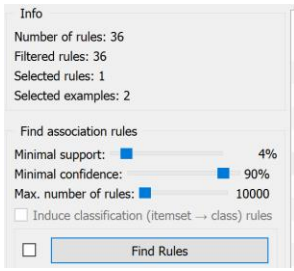
Fig. 2. Frequency distribution of
Purchased Grocery Items without Bread

Following the observations for support, confidence, and lift the focus became the association rules themselves characterized using antecedents and consequents. What means is that if a customer purchases items A and B, the antecedent, then they are likely to buy item C, the consequent, as well. These are our rules as defined by the apriori algorithm.

Before constructing the association rules, the data set was preprocessed by creating an equal frequency discretization using discretize continuous variables. The association rules where then determined by setting a minimum support of four percent along with a minimum confidence of 90%. The results were then organized by highest lift value and are discussed in Section IV.

IV. Results and Discussion

The results yielded from our methodology in Section III resulted in thirty-six rules based on our modified list of ten features. After organizing them by highest lift scores there were five rules that stood out. Ideally I would have liked to show the connections between these five association rules with a Chord Diagram, but Orange does not support this visualization tool. The association rules can be seen instead in Figure 3 and all have a lift score of ten which implies a high likelihood of these five sets appearing in the same transaction compared to their general purchase rates. Ten was the highest lift score which was derived from our apriori algorithm.



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.100	1.000	0.100	1.000	10.000	0.090	BISCUIT=single_value, COFFEE=single_value	→	CORNFLAKES=single_value, APPLE=single_value
0.100	1.000	0.100	1.000	10.000	0.090	CORNFLAKES=single_value, APPLE=single_value	→	BISCUIT=single_value, COFFEE=single_value
0.050	1.000	0.050	2.000	10.000	0.045	MILK=single_value, MAGGI=single_value	→	JAM=single_value
0.050	1.000	0.050	2.000	10.000	0.045	TEA=single_value, COFFEE=single_value	→	MILK=single_value, CORNFLAKES=single_value
0.050	1.000	0.050	2.000	10.000	0.045	MILK=single_value, COFFEE=single_value	→	CORNFLAKES=single_value, TEA=single_value
0.100	1.000	0.100	1.500	6.667	0.085	BISCUIT=single_value, COFFEE=single_value	→	APPLE=single_value
0.100	1.000	0.100	1.500	6.667	0.085	BISCUIT=single_value, CORNFLAKES=single_value, COFFEE=single_value	→	APPLE=single_value
0.100	1.000	0.100	2.000	5.000	0.080	BISCUIT=single_value, APPLE=single_value	→	CORNFLAKES=single_value, COFFEE=single_value
0.050	1.000	0.050	4.000	5.000	0.040	MILK=single_value, TEA=single_value	→	CORNFLAKES=single_value, COFFEE=single_value
0.100	1.000	0.100	2.500	4.000	0.075	BISCUIT=single_value, TEA=single_value	→	MAGGI=single_value

Fig. 3. Strongest Association Rules based on Lift Scores

An interesting example of what was found from these results was the antecedent of MILK and MAGGI and the sets consequent of JAM. The feature JAM has the lowest frequency in Figure 2, appearing in only two of the transactions recorded. The implications of this will be discussed in Section V. These five associations will be the strongest influencers as to what conclusions can be drawn from the results in Figure 3.

V. Conclusions

The results from Section IV can be leveraged by the UK grocery store where these transactions were recorded by placing associations with high lift scores closer to one another to incentivize more sales for grocery products. Referring back to the association in which JAM was the consequent, we see an opportunity to increase the sale of JAM by placing the grocery item alongside MILK and MAGGI. Similarly, the sale of APPLES in this grocery store may be increased by placing the item next to where the BISCUIT and COFFEE are placed, however I would be hesitant to suggest this since after removing BREAD, BISCUIT and COFFEE have become very popular products and may not be surprising that the two items appear in many antecedents.

References

[1] https://www.kaggle.com/shazadudwadia/supermarket