SAINT MARY'S UNIVERSITY SINCE 1802

One University. One World. Yours.

Data and Text Mining

MCDA5580

Team Members:

NAME	A#	Email
Manoj Bandaru	A00433174	Manoj.bandaru@smu.ca
Aditya Tandon	A00432835	Aditya.Tandon@smu.ca
Khagesh Pandya	A00431429	Khagesh.pandya@smu.ca

Table of Contents

1. Executive Summary:

We were given the data set of online retail store and by using association mining algorithms perform market basket analysis to get frequent item set and rule generation so that onlineretail can get their inventory and discounts in place and have better insights about the frequently bought items and their customers. It will find out underlying patterns of customer purchase behavior which will allow businesses make effective decisions.

2. Objective:

Our objective is to use Apriori algorithm which is one of the association mining algorithms. Association mining rule viewed on Frequent Item generation and Rule generation. When you apply Association mining rule on a given set of transactions T your goal will be to find all rules with[1]:

Support greater than or equal to min_support

Confidence greater than or equal to min_confidence

3. Data Summary:

The Data has we have had the below columns:

Invoice No: It is unique transaction ID to identify transaction.

StockCode: Stock Code is unique ID for product.

Quantity: Number of quantities sold for each product in a given transaction.

Description: Description about product.

Invoice Date: Transaction date and time in string.

Unit Price: Unit price for each product.

CustomerID: Unique ID for the given customer.

Country: From where transaction took place.

Invoice Date Time: Timestamp for the transaction.

The Parameters or the columns we used for association mining are CustomerID, Stock Code Description

3.1 Data Cleaning:

Data has 0.5 million transactions, but it needs to be cleaned. So for that we use Microsoft excel to filter out rows. Below are steps which performed on the columns.

InoviceNo: Removed transaction id's which were inappropriate like 0

StockCode: Kept only alphanumeric and numeric values.

Quantity: Removed negative values.

UnitPrice: Removed negative and 0 values.

Description: Removed blanks and other unnecessary values.

CustomerID: Removed 0's from CustomerID as well.

After cleaning the data, we used 159479 rows for association mining.

4. Data Observation:

According to our observation, we got some interesting insights on the dataset. There were a lot of customers who were doing multiple transactions on a single day, which gave us a sense that it could be possible that a single customer maybe isn't buying a single thing, because of some discount on number of items, or same products maybe. We ran a SQL query to find on a single date, a single customer has done multiple transactions or not, and we found that there are customers who are buying products on multiple transactions on same day, we clubbed those products which are bought on same day and ran algorithm on that data. We got some interesting pairs after running the apriori algorithm on that data.

Columns Considered: Description, CustomerID, Date

Data Size: After Cleaning data it contain 7357 rows are there.

Association Base: As discuss above data is transposed based on **CustomerID** and **Date** of transaction.

5. Association Mining:

5.1 Approach:

1. We cleaned the data and used the CustomerID and Date combination to take all the transactions in a date and combined all the items bought by that customer in a single day.
2. Taking the data, after applying the ddply function in R, we trained it with the Apriori algorithm to generate certain sets of rules, where we gave the minimum support level as 0.1% of the total dataset, and the 90% confidence level.
3. Since, we took the support level as low as 0.1 and confidence level of 90%, we sorted the rules by the confidence level 100% and generated some unique itemsets with it.
4. After generating the itemsets, we sorted the itemsets with the support value, to get only those items, which have higher support level. We took the items in the range of 0.6-1.2% of support value.
5. From the itemsets, we generated the maximally frequent itemsets, using the is.maximal function and analyzed the itemsets and gave our analysis.

5.2 Support and Confidence:

Support (s): Fraction of transactions that contain the item-set 'X' [1]

$$\text{Support}(X)=\text{frequency}(X)/ N$$

Confidence (c): For a rule A=>B Confidence shows the percentage in which B is bought with A.

$$\text{Confidence}(A=>B)=P(A\cap B)/P(A)$$

So, according to definition, we realized that it is a dataset of about 0.5 million transactions, so we need to take the minimum support of 0.001, which means about 500 transactions should be containing the same pattern, with a confidence level of at least 90%, which means that 90% of the transaction containing any of these items have all the other items for which the rules[3] have been generated.

5.3 Lift:

Lift gives the correlation between A and B in the rule A=>B. Correlation shows how one item-set A effects the item-set B.[1]

$$\text{Lift}(A=>B)=\text{Support}/\text{Supp}(A)\text{Supp}(B)$$

As the lift gives you the co relation between item sets and lift of 1 is mere co-incidence, negative or less than 1 lift tells you that two items are not bought together, and the negative value is also good for rule generation, more than one lift tells that the co relation between two item sets have dependency.

Rules based on lift:

```
    lhs                                          rhs                          support          confidence    lift
[1]   {SMALL POP BOX}                => {FUNKY MONKEY}               0.001072907  1.0000000 932.0476
[2]   {FUNKY MONKEY}                  => {SMALL POP BOX}              0.001072907  1.0000000 932.0476
[3]   {TREES}                     => {CHRISTMAS GARLAND STARS}          0.001328360  1.0000000 752.8077
[4]   {CHRISTMAS GARLAND STARS}           => {TREES}                    0.001328360  1.0000000 752.8077
[5]   {BILLBOARD FONTS DESIGN}           => {WRAP}                    0.001481633  1.0000000 631.3871
[6]   {WRAP}                       => {BILLBOARD FONTS DESIGN}           0.001481633  0.9354839 631.3871
[7]   {MIRRORED WALL ART LADIES}  => {MIRRORED WALL ART GENTS}0.001123997  0.8461538 534.250
[8]   {FUNK MONKEY}                  => {ART LIGHTS}                    0.002196904  1.0000000 455.18
```

5.4 Frequent Item Set:

Finding all frequent item-sets with support >= pre-determined min_support count. We generated Frequent Item sets based on the support which is greater than pre-determined min_support[4].

Below are the frequent Item sets:

```
                    items                                support
[1]  {SET 3 RETROSPOT TEA,SUGAR}               0.015357434
[2]  {COFFEE,SUGAR}                            0.015357434
[3]  {COFFEE,SET 3 RETROSPOT TEA}              0.015357434
[4]  {BACK DOOR,KEY FOB}                       0.012367491
[5]  {GARAGE DESIGN,KEY FOB}                   0.011144333
[6]  {BIRTHDAY CARD,ELEPHANT}                  0.010192987
[7]  {FRONT  DOOR,KEY FOB}                     0.007338951
[8]  {DECORATION,METAL}                        0.006795325
[9]  {COFFEE,SUGAR JARS}                       0.006251699
[10] {1 HANGER,HOOK}                           0.006251699
[11] {HOOK,MAGIC GARDEN}                       0.006251699
[12] {1 HANGER,MAGIC GARDEN}                   0.006251699
[13] {BREAKFAST IN BED,TRAY}                   0.006115792
[14] {BIRTHDAY CARD,RETRO SPOT}                0.006115792
[15] {CUPCAKE SINGLE HOOK,METAL SIGN}          0.005300353
[16] {AIRLINE LOUNGE,METAL SIGN}               0.005300353
[17] {DECORATION,WOBBLY RABBIT}                0.004620821
[18] {METAL,WOBBLY RABBIT}                     0.004620821
[19] {SUGAR JARS,WHITE TEA}                    0.004077195
[20] {COFFEE,WHITE TEA}                        0.004077195
```

5.5 Maximal Frequent Sets:

An itemset is maximal in a set if no proper superset of the itemset is contained in the set. We define here maximal rules, as the rules generated by maximal item sets with a function is.maximal().[6]

Items	Support
[1] {SET 3 RETROSPOT TEA,SUGAR}	0.015359522
[2] {COFFEE, SUGAR}	0.015359522
[3] {COFFEE, SET 3 RETROSPOT TEA}	0.015359522
[4] {KEY FOB, SHED}	0.014000272
[5] {BACK DOOR, KEY FOB}	0.012369172
[6] {GARAGE DESIGN, KEY FOB}	0.011145847
[7] {BIRTHDAY CARD,ELEPHANT}	0.010194373
[8] {FRONT DOOR, KEY FOB}	0.007339948
[9] {DECORATION, METAL}	0.006796248
[10] {1 HANGER, HOOK}	0.006252549
[11] {1 HANGER, MAGIC GARDEN}	0.006252549
[12] {HOOK, MAGIC GARDEN}	0.006252549
[13] {BREAKFAST IN BED, TRAY}	0.006116624
[14] {BIRTHDAY CARD, RETRO SPOT}	0.006116624
[15] {CUPCAKE SINGLE HOOK, METAL SIGN}	0.005301074
[16] {AIRLINE LOUNGE, METAL SIGN}	0.005301074
[17] {DECORATION, WOBBLY RABBIT}	0.004621449
[18] {METAL, WOBBLY RABBIT}	0.004621449
[19] {BILLBOARD FONTS DESIGN,WRAP}	0.003941824
[20] {DECORATION, WOBBLY CHICKEN}	0.003805899

5.6 Rules Generated:

Rule Generation is to list all Association Rules from frequent item-sets. By calculating Support and Confidence for all rules. Pruning rules that fail min_support and min_confidence thresholds.[1]

```
                 lhs                                   rhs                          support      confidence     lift
[1]   {GLITTER CHRISTMAS STAR}     => {GLITTER CHRISTMAS HEART}  0.001087252          1        459.87500
[2]   {NEW ENGLAND}                  => {TUMBLER}                    0.001494971           1       525.57143
[3]   {ELVIS LIVES}               => {S/4 ICON COASTER}              0.001223158           1        817.55556
[4]   {S/4 ICON COASTER}              => {ELVIS LIVES}               0.001223158           1       817.55556
[5]   {FUNKY MONKEY}                 => {SMALL POP BOX}                 0.001494971         1       668.90909
[6]   {SMALL POP BOX}                => {FUNKY MONKEY}                  0.001494971         1       668.90909
[7]   {PINK  SPOTS}                 => {SWISS ROLL TOWEL}              0.002310410          1        229.93750
[8]   {CAROUSEL PONIES BABY BIB}=> {SCOTTIE DOGS BABY BIB}       0.001359065          1       319.91 304
[9]   {BLACK TEA}                   => {SUGAR JARS}                 0.003533569              1        159.95652
[10]  {BLACK TEA}                   => {COFFEE}                  0.003533569                1       46.86624
[11]  {ART LIGHTS}                  => {FUNK MONKEY}                0.002174504             1      459.87500
[12]  {FUNK MONKEY}                   => {ART LIGHTS}               0.002174504             1      459.87500
[13]  {DECOUPAGE}                    => {GREETING CARD}               0.002718130           1      210.22857
[14]  {WOBBLY CHICKEN}                 => {DECORATION}                  0.003805382         1      147.16000
[15]  {WOBBLY CHICKEN}                 => {METAL}                 0.003805382               1      147.16000
[16]  {LIGHT PINK}                  => {FEATHER PEN}               0.002989943              1      133.78182
[17]  {CHOCOLATE  SPOTS}                => {SWISS ROLL TOWEL}               0.003669475       1      229.93750
[18]  {HOT PINK}                    => {FEATHER PEN}               0.003805382              1     133.78182
[19]  {NURSERY A}                    => {B}                   0.003805382                  1     262.78571
[20]  {B}                        => {NURSERY A}                0.003805382                    1     262.78571
```

6. Conclusion:

- As we ran Apriori algorithm to get better insights of the customer behavior and basket analyzing, we found pretty good insights for OnlineRetail.
- As we approached Our analysis based on Customer who is buying the items on different Invoice No but on same date, we got good rules which lead to get better Frequent Item sets in the basket.
- We also got good lift values which state the better co relation between the item sets in the rule.
- As per our insights and to justify that we got good frequent Item sets we are mentioning few of the frequent Item sets with support value greater than defined support value.

{SET 3 RETROSPOT TEA, SUGAR} -> If someone is buying tea, they would buy sugar too.

{COFFEE, SUGAR} -> If you are buying coffee, you would take sugar too.

{COFFEE, SET 3 RETROSPOT TEA} -> Someone who buys tea, will also buy coffee.

{GARAGE DESIGN,KEY FOB} -> To open a Garage, people buy the garage design to decorate the new garage, also, they use the key fob to open the Garage Door.

{BIRTHDAY CARD,ELEPHANT}-> A person buying a birthday card, is planning to gift an elephant toy too.

{FRONT DOOR,KEY FOB} -> When someone buys a front door, they buy a key fob with it.

{1 HANGER,HOOK} -> To hang a hanger, they need a hook, so these two go together.

{HOOK,MAGIC GARDEN} -> If you are setting up the magic garden, you need a hook to put a swing on it, and hence they go together.

{BREAKFAST IN BED,TRAY} -> To serve the breakfast in bed, one buys Tray with it.
{DECORATION,WOBBLY RABBIT} -> If you are organizing a party, then one buys the decoration, and when one buys a decoration, they buy a gift, and as a gift, they buy a wobbly rabbit.

7. References:

Jabeen, H. (2018). *Market Basket Analysis using R*. [online] DataCamp Community. Available at: https://www.datacamp.com/community/tutorials/market-basket-analysis-r [Accessed 21 Jun. 2019][1].

Rdocumentation.org. (n.d.). *sort function | R Documentation*. [online] Available at: https://www.rdocumentation.org/packages/base/versions/3.6.0/topics/sort [Accessed 21 Jun. 2019][2].

Rdocumentation.org. (n.d.). *apriori function | R Documentation*. [online] Available at: https://www.rdocumentation.org/packages/arules/versions/1.6-3/topics/apriori [Accessed 18 Jun. 2019][3].

Rdocumentation.org. (n.d.). *ddply function | R Documentation*. [online] Available at: https://www.rdocumentation.org/packages/plyr/versions/1.8.4/topics/ddply [Accessed 18 Jun. 2019][4].

Li, S. (2017). *A Gentle Introduction on Market Basket Analysis — Association Rules*. [online] R-bloggers. Available at: https://www.r-bloggers.com/a-gentle-introduction-on-market-basket-analysis%E2%80%8A-%E2%80%8Aassociation-rules/ [Accessed 21 Jun. 2019][5].

Rdrr.io. (n.d.). *is.maximal: Find Maximal Itemsets in arules: Mining Association Rules and Frequent Itemsets*. [online] Available at: https://rdrr.io/cran/arules/man/is.maximal.html [Accessed 21 Jun. 2019][6].