Assessment Submission Cover Sheet

This Assessment Cover Sheet **must** be included on all Assessment submissions.

|  |  |
| --- | --- |
| Assignment Title | CA2 – Association Rules |
| Module | Data Mining |
| Student Name  (same as Student Card) | Ciaran Finnegan |
| Student Number | D21124026 |
| Programme |  |
| Part-Time/Full-Time |  |
| Year of Study  (First Year, Second Year, etc) |  |

Late Submissions: Assessment submitted after the deadline will have a late penalty applied.

**Academic Integrity for assessment in TU Dublin Programmes**

Each student is responsible for knowing and abiding by TU Dublin Academic Regulations and Policies. Any student in breach of these regulation/policies will be subject to action in accordance with the University’s procedures for breaches of assessment regulations. Please refer to the General Assessment Regulations at

<https://tudublin.libguides.com/c.php?g=674049&p=4794713>

<https://www.tudublinsu.ie/advice/exams/breachesofregulations/>

All students are expected to complete their courses/programmes in compliance with University regulations. No student shall engage in any activity that involves attempting to receive a grade by means other than honest effort, for example:

1. No student shall complete, in part or in total, any examination or assessment for another person.
2. No student shall knowingly allow any examination or assessment to be completed, in part or in total, for themselves by another person.
3. No student shall plagiarise or copy the work of another and submit it as their own work.
4. No student shall falsify any data. Falsification is the invention of data, its alteration, its copying from any other source, or otherwise obtaining it by unfair means, or inventing quotations and/or references.
5. No student shall use aids or devices excluded by the lecturer in undertaking course work or assessments/ examinations.
6. No student shall knowingly procure, provide, or accept any materials that contain questions or answers to any examination or assessment to be given at a subsequent time.
7. No student shall provide their assignments, in part or in total, to any other student in current or future classes of this module/ programme unless authorised to do so by the lecturer.
8. No student shall submit substantially the same material in more than one module/programme without prior authorization.
9. No student shall alter graded assignments or examinations and then resubmit them for regrading, unless specifically authorised to do so by the lecturer.
10. All programming code and documentation, unless correctly referenced, submitted for assessment or existing in the student’s computer accounts must be the students’ original work or material specifically authorized by the lecturer.
11. Collaborating with other students to develop, complete or correct course work is limited to activities explicitly authorized by the lecturer.
12. For all group assignments, each member of the group is responsible for the academic integrity of the entire submission. Consequently, all group members must satisfy themselves that all elements of their submission adhere to the academic integrity statement points above.

By submitting coursework, either physically or electronically, you are confirming that it is your own work (or, in the case of a group submission, that it is the result of joint work undertaken by members of the group that you represent) and that you have read and understand the University’s Regulations and Policies covering Academic Integrity (see General Assessment Regulations)*.*

Coursework may be submitted to an electronic detection system in order to help ascertain if any plagiarised material is present. If you have queries about what constitutes plagiarism, please speak to your lecturer.

|  |  |
| --- | --- |
| Student Signature |  |
| Date |  |

IMPORTANT:

* Complete the required number of tasks as defined in Assessment Handout
* The sections listed below are an example of the section headings for each task. You can use alternative headings
* Tasks 1-3: Sub-Sections 1-7 should be no longer than 8 pages (minimum 6 pages), including diagrams, images, screen captures, tables, etc. Careful selection of these is needed.
  + Code does not count to this total. Code should be added to the relevant section.
* Detailed discussion is expected. Marks are awarded based on depth of information given.
* Marks are awarded based on complexity of problem and depth of work.

# TASK 2 - *Association Rules – Grocery Market Basket Analysis*

1. **Definition of Problem**

This part of the assignment looks at the concept of association rules being used in market basket analysis.

It begins with a Kaggle sourced dataset (located [here](https://www.kaggle.com/heeraldedhia/groceries-dataset)) of purchase orders by customers in a grocery store. The objective of this exercise is to use SAS Enterprise Miner (EM) to analyse the grocery order dataset and determine what types of associations and correlations can be found between the transactions/products.

The benefits of such an analysis are to determine the strength of the buying patterns and how/if this should influence the layout and presentation of goods in our hypothetical retail grocery store.

Is there insight in this process that can help both the customer and retailer shop faster and smarter? If certain items, or groups of items, are frequently bought together then there is a solid logic in placing these items closer together physically. For an e-commerce experience, it might mean reducing the number of clicks to move between items that are highly related in terms of purchasing patterns.

1. **Data Exploration & Descriptive Analytics**

This Kaggle grocery transaction dataset is relatively straightforward in structure. It contains **3** columns and **38,765** rows.

After selecting this dataset, the first action was to load it into SAS Studio and take a look at how the CSV file was imported, and data types assigned.

Graphical user interface, table

Description automatically generated

Fig 1 – Grocery Dataset Imported Into SAS Studio.

Each row represents a purchase of one item on a given day by a customer.

The layout of our dataset;

* ***Member\_Number***: Unique number ID for customer.
* ***Date***: The purchase date of the transaction.
* ***ItemDescription***: Text detail of the single grocery item itself.

The 38,765 rows in the dataset each represent a unique transaction on one item, but many items are bought together by customers on a given day.

To better understand the information Kaggle provides an overview that in the dataset there are;

* **728** unique dates when grocery shopping took place.
* **167** unique items, which are sold by the shop.

A number of 3898 unique customers can be determined by a quick check on the unique number of customer IDs in the dataset. Below is a code snippet after the dataset was loaded into a Google Colab Python Notebook;

Graphical user interface, text, application

Description automatically generated

Fig 2 – Number of Unique Customers in Grocery Dataset

After loading the data into SAS Enterprise Miner (EM), a quick check on the new data source shows that there are no missing values.

Table

Description automatically generated

Fig 3 – Check on Imported Data for Grocery Data Source in SAS EM

1. **Data Preparation**

Importing the Grocery ***csv*** file into SAS Studio converted the attributes into the correct data type.

It was then relatively straightforward to transfer the data as a Data Source into the project library in Enterprise Miner (EM) and, as Section 2 of this report section shows, the data was fully intact (no missing attributes).

To execute an Association Rules analysis some changes are required in SAS EM to the Data Source, and its attributes.

The Data Source ‘role’ was changed to ‘Transaction’;

Graphical user interface, text, application, email

Description automatically generated

Fig 4 – Change Grocery Data Source Role to ‘Transaction’

During the importing of the Data Source into the SAS EM project, the attributes in the Grocery dataset are also altered.

Table

Description automatically generated

Fig 5 – Update Column Metadata in Grocery Data Source

The column for ‘itemDescription’ contains the element for which we are trying to establish patterns with our Association Rules analysis, hence it is assigned the *Role* of ‘Target’.

1. **Details of Algorithms & Configurations**

A diagram is created in SAS EM and the amended Data Source with the Grocery transaction information is added.

An ‘Association’ node is then connected to perform the actual association rules analysis.

Shape

Description automatically generated with low confidence

Fig 6 - Connecting Association Node to Data Source.

Most of the default settings in the node can be left unchanged, based primarily on guidance from DM lecture lab notes. The ‘Export Rule by ID’ is changed and set to ‘Yes’;

Table

Description automatically generated with low confidence

Fig 7 – Change Export ID Rule

This allows the actual Association Rules to be saved. The Rules Description table will be made available in the ‘Results’ window display for review.

The last configuration change is to select the ‘Variables’ option under the ‘Train’ set of settings and update so that the ‘Date’ attribute is not used in the subsequent analysis.

Table

Description automatically generated

Table

Description automatically generated

Fig 8 – Set ‘Date’ Attributes To ‘No’

1. **Model Performance Metrics & Evaluation of Results**

After the ‘Association Node’ is run we can assess the ‘Results’.

*Rules Tables*

The first set of Result to assess are the ‘Rules Table’ and the ‘Output’, which is essentially the same data in different formats.

Table

Description automatically generated with medium confidence

Fig 9 – Rule Table

Table

Description automatically generated

Fig 10 – Results Output Table

Both of these diagrams are the Association Rules that shows the strongest links between grocery store items or groups of items. The rules are in descending order of ‘Lift’. (Only the very highest entries are displayed here).

The ‘*Support*’ value shows the popularity of a grocery item (or a group of items) in terms of the numbers of customer transaction that selected that item(s).

‘*Confidence*’ represents the likelihood of the RHS (Y) item being purchased if the LHS (x) has already been bought. Thus, from the above table, there is a 20.5% chance of *whole milk and sausage* being bought if the customer has already purchased *yogurt and rolls/buns*. The problem with such a metric is that very popular items can distort the confidence value and suggest a correlation that is stronger than it actually is.

‘Lift’ attempts to counteract the problems with using the ‘Confidence’ measure by adjusting the probability to counter for very popular RHS (Y) items. It can thus be a more accurate reflection of the correlation between items, or groups of items, and this is why the above tables show a scale based on ‘Lift’ values.

If the Lift value is above 1 then there is an increasing association. The closer to 1 the lower the association, and values lower than 1 indicate a negative correlation.

The ‘*yogurt & rolls/buns ==> whole milk & sausage*’ is the strongest association in the dataset and is followed in the diagrams above by its reciprocal rule.

*LHS v RHS*

The LHS – RHS Rule Matrix contains a great deal of data but is colour coded to help with ad-hoc analysis.

Taking a quick example of Yogurt as a LHS Value;

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

Fig 11 - Yogurt Confidence

As one moves along a straight line from left to right, the changing degrees of confidence that RHS items have in relation to Yogurt (LHS) can be shown. All these cells are pale blue indicating a lower degree of confidence.

Looking at Whole Milk, as one moves from bottom to top the confidence values for Whole Milk to other items can be seen. These are darker red dots and of much more interest to the store owner in terms of strong product correlation.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Fig 12 – Whole Milk Confidence

One further output of interest from SAS EM is a Link Graph to visually show the strength of relationships between the store items. This next diagram is somewhat hard to read but emphasises the analysis from the other SAS EM Result outputs.

A picture containing indoor, map, colorful

Description automatically generated

Fig 13 - Link Graph output from SAS EM

Overall Observations from Association Rules Analysis on Grocery Dataset

* *Whole Milk* is the item with the strongest associations with other items. It is represented by the deepest red dot in the Link Graph and tops the Rules Tables. Not surprisingly it associates strongly with items such as vegetables and yogurt.
* Various vegetable types, yogurt, rolls/buns, and soda all feature with strong relationships to other items.
* Soda is perhaps the least ‘healthy’ of the main items, but the strength of the associations might warrant a review of its location on shelves within the physical retail outlet.

*Adding a Chi Square Statistic*

A SAS EM research paper in 2012**[1]** proposed adding a SAS Code Utility node with Chi Squared analysis code to a Grocery Market Basket analysis. The purpose was to provide an additional Results output to show if any of the association rules were not statistically significant, and hence if the overall set of rules could be reduced to a smaller size.

The source code was taken from the appendix of that research paper and added to our Associations Rules diagram in SAS EM.

Graphical user interface, text, chat or text message

Description automatically generated

Fig 14 - Adding SAS Code Node (which has been renamed) to Diagram.

Although this was an interesting exercise, there were no rules in the 200-rule set in this experiment with a high enough *PValue* to declare them as not statistically significant (at alpha = .05).

Graphical user interface, text, application

Description automatically generated

Fig 15 – Output From Chi Squared Analysis (excerpt of top ten rules)

1. **Comparison with Other Research**

*Enhance the decision-making process around business rules*

A paper published by Faron and Chakraborty at the SAS Global Forum in 2012**[1]** suggested a way to add the Pearson’s Chi Squared statistic to the output from the SAS EM *Association* node.

The addition of this test would enhance the quality of the data analysis by showing which rules are statistically significant, therefore adding an additional metric the evaluate the importance of given association rules.

**I implemented this Chi Squared analysis, but it did not yield any obvious benefits (or none which I could determine).**

*Representing Rules: Visual Clutter*

The Link Graph output in SAS EM is a useful at a glance tool to visually represent the associations rules in our grocery database. However, research on similar grocery datasets, such as that by Hahsler, M., Hornik, K. and Reutterer, T., 2016**[2]**, highlights the challenge that these graphs easily become cluttered as the number of rules grow. (This would be true of any rule set, not just the typical grocery basket analysis being conducted in this section of the assignment).

If a visual representation of the association rules is an important outcome, then there are tools such as ***arulesViz*** that offer more sophisticated interactive functionality. For the purposes of this assignment the SAS EM graph tools were deemed sufficient.

*What to do with these Association Rules?*

This dataset is a relatively straightforward representation of grocery purchases, and the common benefit is often considered to be a more effective physical retail shop layout.

A key point made in a 2018 Towards Data Science article**[3]** is that Association Rules look at lists of items with unique transaction ID from many users, and studies these lists as a block. This is not an approach that generates a recommendation for one *specific* user. That said, research I found on the role of Association Data Mining and E-Commerce website structure**[4]** shows how the selection of *antecedent* lists can be used to meaningfully direct a single user to different web pages, where they are most likely to find the products for which they are looking.

A follow-on challenge for me would be to repeat this exercise with a similar but extended dataset that looked at product groupings in the *itemsets*, and also considered user profiles.

1. **References**

[1] Faron, M. and Chakraborty, G., 2012. *Easily Add Significance Testing to your Market Basket Analysis in SAS® Enterprise Miner*. [online] Citeseerx.ist.psu.edu. Available at: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.421.1785&rep=rep1&type=pdf> [Accessed 5 January 2022].

[2] Hahsler, M., Hornik, K. and Reutterer, T., 2016. Implications of Probabilistic Data Modelling for Mining Association Rules. *From Data and Information Analysis to Knowledge Engineering*, [online] pp.598-605. Available at: <https://link.springer.com/content/pdf/10.1007/s11573-016-0822-8.pdf> [Accessed 5 January 2022].

[3] Garg, A., 2018. *Complete guide to Association Rules (1/2)*. [online] TowardsDataScience. Available at: <https://towardsdatascience.com/association-rules-2-aa9a77241654> [Accessed 31 December 2021].

[4] Omari, A., Conrad, S. and Alcic, S., 2007. Designing a Well-Structured E-Shop Using Association Rule Mining. *2007 Innovations in Information Technologies (IIT)*, [online] Available at: <https://ieeexplore.ieee.org/abstract/document/4430429> [Accessed 5 January 2022].