STAT828 Data Mining – Undirected Knowledge Discovery (MBA) Project

Due date: Tuesday, 9 April 2019, 9:00 am

This is an individual project. If you work with another student, you need to acknowledge it in your report.

Prepare a report for the Simulated Coles data set. This data set contains variables similar to you would expect in a transactional data set (i.e. shopping basket) and demographic variables about the customers. This data set is created by A/Prof Ayse Bilgin.
If you have your own data, you may be able to use it subject to the agreement of A/Prof Ayse Bilgin.

The aims of this project are

- to explore the data set in IBM SPSS Modeler or R or any other software, then do data preprocessing
- to apply cluster analysis techniques to discover clusters in a data set in IBM SPSS Modeler and/or R;
- to apply Market Basket Analysis techniques to discover patterns in the data set in IBM SPSS Modeler and/or R;
- to create association rules in SPSS Modeler and/or R;
- to write (a part of) a professional report of your findings including the data description based on your explorations; and
- to assess your own learning.

In summary your report should include the following: State the Business Problem for your report. Explore the dataset to understand what is in it (descriptive data mining). You can use visualization tools from R and/or IBM SPSS Modeler or other software packages. When you know what is in your dataset, then you can decide which fields (variables) will be used in your data mining explorations, accordingly you can clean and/or prepare your data (i.e. creating concept hierarchies, transforming some of your variables or creating new variables from the ones in the data set) for the predictive data mining. Apply suitable model(s) to your data and present your results. When applying cluster analysis to discover clusters/patterns in the data set, you need to describe the created clusters and compare clustering results from different methods to see which one is easier to interpret. Try to label the created clusters before you interpret them. Split your results section into two parts, one is the Cluster analysis results, two is the Market Basket Analysis results. You need to explain both results, just presenting computer output is not sufficient. Finally, describe how your findings could be used in day-to-day operations of the company.

What to submit:

1. Upload your project report on iLearn by due date (as word document or pdf) clearly stating your name and your student ID both in the document and the document name.
2. Upload related files on iLearn, such as R codes (as an R script file or text file) and IBM SPSS Modeler streams used for your analysis.
3. Answer the "Feedback on the Project" questions and upload it either as the last page of your report or as a separate word document.

Feedback on the Project (Self Assessment)

1. Which part of this project you struggled most?
2. Did you discuss your difficulties with your peers?
3. What would have been helpful to overcome your difficulties identified in Q1?
4. Was the time given for this project sufficient?
5. Did the feedback you received for the lab exercises help you with this project?
6. Did the reading materials provided up till now help you with this project?
7. Reflect on one of the employability skills you gained by completing this project (no more than 250 words).

The Report Template

Use Microsoft Word or a similar open version of word or pdf to create a Project Report, containing the following sections (see marking scheme for fine details):

- **Executive Summary or Abstract:** Give the big picture first and then in two or three sentences summarise the essence of what you have done *and discovered* in no more than 15 lines (approximately 250 words). It is a bit like telling the punch line of a joke before telling the joke, but busy executives insist on it these days – they don't want to be distracted by all the whys and wherefores, they just want to "cut to the chase" and get a distillation of what has been achieved.
- **Introduction**: This section provides the questions to be answered.
  Pose two questions to be answered, one for the cluster analysis and one for the market basket analysis. (Imagine yourselves being the manager of the company that its data being analysed while you are writing your questions and your report.)
- **Description of the data set:** This section is for data understanding and descriptive data mining
  - **Original Data:** List the number of attributes (variables), number of observations (rows), and format of the data, followed by a description of your data set (i.e. graphs, tables).
  - **Data preprocessing:** Explain how you preprocessed your data. Did you find any outliers, if so what did you do? How about missing values? Did you create any new variable(s) from existing variables, if so, how and why? Did you exclude any variables or observations? Who are the customers? What was the most/least purchased item?
- **Methodology:** This section gives the details of modeling methods.
  - Briefly describe the modeling methods you have used. Explain the reasons for using a specific method and not any other.
  - You also need to explain the reasons for choosing a particular modeling option for each specific model.
  - **Data mining evaluation (optional):** You might create two sets of association rules by randomly splitting the data set into two files and then compare the consistency of the rules in two outputs. This is not a bullet proof goodness check since association rules are part of unsupervised learning tools and we do not have an outcome variable to assess the developed models. However consistency of the rules for the two sets of data could be used as an indication of the good model. Similarly, you can compare clusters created by subsets of the data set.
- **Results:** Display and explain the results, if possible, in layman terms. For the questions identified above, state the knowledge you have discovered that provides insights into this question(s). State how well the knowledge answers the question(s), what is missing or requires further analysis.
- **Conclusion:** Provide an ending to this document with a mention of how your findings could be used in day-to-day operations and identify possible extensions for the future. In addition, identify the limitations of your data set and analysis, if there were any. Your conclusion may have quite a lot in common with the "Executive Summary" but it should be more detailed. (one paragraph, roughly 250 words).
- **Appendix:** Computer outputs for all the models applied should be presented here. In addition, anything that is important to present but will take too much space in the main report should be placed in the appendix. Make sure you refer to everything in the appendix in your report.

Note: Your main report should not be more than 8 A4 pages. All relevant graphical displays should be embedded in your report. Font type for the report is Times New Roman. Font size is 12 in the body. Line space is 1 or 1.5.

This assignment is designed to help you develop the following graduate capabilities:

1. Discipline specific knowledge and skills (identify and apply appropriate data-mining techniques to a new problem; demonstrate the use of the freeware package "R" and/or SPSS Modeler in carrying out some of these data-mining techniques)
2. Critical, analytical and creative thinking (identify and embed a graphical display when it is appropriate; identify the suitable data pre-processing methods for a new problem; and apply a suitable data mining technique to current problem; examine and compare the differences between different models and interpret the results from sophisticated models to end users. In addition, discuss the limitations of such a method)
3. Problem solving and research capability & Creative and innovative (identify which variables will be cleaned, transformed and apply suitable methods to deal with these and similar problems (i.e. outliers, missing data); identify a business problem for a data set and propose a solution for the problem by applying suitable data mining methods)
4. Effective communication (written communication using formal language and academic conventions, particularly report writing skills)
5. Socially and environmentally active and responsible (prepare a professional report for the results of the analysis where an action plan for the management is outlined)
6. Professional and personal judgment and initiative (analysis, discussion and conclusions; and self-assessment)
7. Commitment to continuous learning (identify R packages or similar free software that could be used for analysis and then learn them yourself to be effectively solving the problem in hand; share the learning experience with peers in class)
8. Engaged and ethical local and global citizens (when making decisions to include variables for analysis, consider the implications for the local or the global society; report any outliers that were excluded from the study and explain why)

The Criteria for the Project Report

HD: There is substantial originality and insight in identifying, generating and communicating competing arguments, perspectives or problem solving approaches; critical evaluation of problems, their solutions and their implications; creativity in application as appropriate to the discipline.

D: There is demonstration of frequent originality in defining and analysing issues or problems and providing solutions; and the use of means of communication appropriate to the discipline and the audience.

Cr: There is demonstration of substantial understanding of fundamental concepts in the field of study and the ability to apply these concepts in a variety of contexts; convincing argumentation with appropriate coherent justification; communication of ideas fluently and clearly in terms of the conventions of the discipline.

P: There is demonstration of understanding and application of fundamental concepts of the field of study; routine argumentation with acceptable justification; communication of information and ideas adequately in terms of the conventions of the discipline. The learning attainment is considered satisfactory or adequate or competent or capable in relation to the specified outcomes.

F: There is missing or partial or superficial or faulty understanding and application of the fundamental concepts in the field of study; missing, undeveloped, inappropriate or confusing argumentation; incomplete, confusing or lacking communication of ideas in ways that give little attention to the conventions of the discipline.

The Marks for each part of the Project Report

Executive Summary:	/5
Introduction:	/5
Description of the data set:	/10
Methodology:	/10
Results:	/10
Conclusion:	/5
Other:	/5
Total Mark	/50

The Marks for each part of the Project Report **Student ID or Name:**

Part of Report	Details covered	Mark
Executive Summary or Abstract	a) Sets the scene (background) b) Specifies what has been done (methodology) c) Summarises results or presents most important results d) Informative	/5
Introduction	a) One question for cluster analysis b) One question for market basket analysis	/5
Description of the data set	a) Errors are identified and if possible, corrected in the data set b) Outliers are identified, if any, and a solution suggested c) Missing data is identified, if any, and a solution suggested d) If new variables are created, the reason(s) and how they were created are explained e) If concept hierarchies are created the reason for creation is explained f) If any observation is excluded from analysis, the reasons are provided g) If any variable is excluded from analysis, the reasons are provided h) Descriptive data mining is performed (i.e. summary statistics for variables and/or graphical displays) and a summary is written i) The most and least common products are identified j) An answer is given to "who are the customers?" question	/10
Methodology	a) The suitability of the methods used for the analysis are explained b) The options chosen to create the clustering models are specified (can be replicated by another researcher) c) The options chosen to create the association rules (i.e. thresholds for support, confidence) are specified (can be replicated by another researcher) d) The reasons for chosen options/thresholds are provided e) (optional) Data mining evaluation is done by creating clusters and/or rules in subsets of the data set	/10
Results	a) The results for cluster analysis are displayed graphically and how they answer the question(s) explained b) The (chosen) association rules are displayed and how they answer the question(s) discussed c) The association rules are explained correctly d) The support of the rules are explained correctly in layman terms e) The confidence of the rules are explained correctly in layman terms f) The lift of the rules are explained correctly in layman terms	/10
Conclusion	a) How the key findings could be used in day-to-day operations is suggested b) The limitations of the data set and/or analysis (if there were any) are discussed c) Future research is suggested	/5
Other Report characteristics	a) Pages are numbered b) Report has no more than 8 pages (excluding appendices) c) All tables and graphics have informative captions d) All tables and graphics are explained and referred in the text e) No (minor/major) spelling/grammatical mistakes f) Computer outputs or R codes are used sparingly in the body of the report g) R script and/or IBM SPSS Modeler stream and/or IBM SPSS Statistics output files (.spv) are submitted h) Feedback to Project is submitted	/5
	Total Mark	/50