COMP5121 Data Mining & Data Warehousing Applications
Assignment 2 Part III

This part of the Assignment is designed to let you experience and go through the steps in Knowledge Discovery in Databases (KDD) process and apply what you have learnt in this subject to work on a practical case and dataset. You have been given a scenario and a dataset and you need to set your own objective(s) to perform your KDD process and data analysis tasks. [Please refer to the later sections for more information and details. Description of Scenario and Dataset are given in the last page of this document]

Requirements

This is a group-based assessment. You are required to form a group of NOT more than FOUR (4) members, i.e. a maximum of FOUR (4) members to work on this group project. Please allocate among yourselves the tasks and the contributions to each task.

Please write the documents in your own words and make sure the materials used have been properly referenced. There is no upper or lower limit for the references in your work.

You are free to choose and use any Data Analytic/Mining Tools, e.g. Orange Version 3, Weka, or you could use any programming languages and packages to help you perform any data analysis and visualization tasks

Every group need to perform a presentation to present, explain and discuss their findings. Every group also need to participate in the briefing session, which is held before all the presentation sessions. Please indicate your assigned Group ID on every deliverables handed-in.

Project Submission Deadline: 23 JUL 2017 (SUN) 12:00 (noon)

Deliverables

- A set of presentation slides (30 slides or less) that will be used to present the suggestions, results and any useful information, as a result of the analysis on the given operational business data, in a 15 minutes presentation. You are reminded that you are NOT asked to write a report (You will lose marks in Quality of the Presentation Slides aspects if you cramp too much text or information into one single slide). You should use diagrams, whenever it is appropriate, to help you illustrate your ideas or visualizing the data to be presented.

  You are reminded that you are going to present to the senior management of your company (the audiences). You should not assume the audiences of the presentation have ANY technical knowledge in KDD or Data Mining. Thus, you are advised to think and speak using the terms in the domain that the audiences can understand in order to present, explain and discuss your findings.

  You are FREE to choose ANY presentation style and flow for preparing the set of slides.

The presentation may contain, but not limited to the following items (This list is for your reference only and if you have a better organization, you are encouraged use your own):

- #Objectives of the KDD Process, e.g.
  - Identifying Problems
  - Identifying Possible Causes/Sources
  - Providing Suggestions/Proposed solutions
- Data Analysis/KDD Methods and Steps
  - Data Cleaning, Preprocessing, ETL, etc.;
  - Data Analysis;
  - Data/Result Evaluation;
- Results of the KDD Process
- Discussion on the KDD Process, e.g.
  - Suggestion(s), Recommendation(s), …
- Conclusion
- References
- *Contributions of Team Members
- *Checklist of performed/discussed steps in the KDD Process

This is the core and important item that guides you perform all the steps in the KDD process and data analysis tasks.

* You must provide us these items in your presentation slides while you may not need to present these items during your presentation.

Submission

◆ A set of presentation slides

You are required to submit ONLY the softcopy of your presentation slides for your group project to LEARN@PolyU.

◆ [Optional] Supplementary Files

If you have other supplementary files that you would like to submit, you MUST include all your files to be submitted into a SINGLE archive, e.g. a ZIP file. Please use the following filename convention for naming the filename of your archive to be submitted:
COMP5121_<your assigned group number>.zip, e.g. COMP5121_EG00.zip

Please ask one of your group member to email the URL of the properly named SINGLE archive file, which contains all your supplementary files to be submitted, for downloading while carbon copy (c.c.) to all your group members.

Please do not attach the SINGLE archive file as email attachment for submission

Please use the following title for your email submission:
COMP5121 <your assigned group number> Project Supplementary Files Submission, e.g. COMP5121 EG00 Project Supplementary Files Submission

Note:

- Please test the validity of all the files and archive to be submitted and, if any, the URL in email submission of supplementary files, for downloading the archive, before you make your submission.
- Your group project submission will not be acknowledged. If there is any problem with your submission, your group will be contacted individually.

Do not submit ANY file of your group project as email attachment.

Do not submit ANY hardcopy of group project report to us.

Late submission of any of the above items may be penalized.

Presentation

Please reserve a time slot, for your presentation, on 2017 JUL 23 (SUN) afternoon using the Discussion (Forum) facility on LEARN@PolyU, "Project Presentation Session Registration", which has been setup already.

You should have submitted the deliverables before conducting your presentation.

Assessment

Your group project will be assessed based on the following criteria

- The Values and Objectives of the Data Analysis;
- The appropriateness and quality of how the steps in KDD Process are performed;
- Quality of the Presentation Slides (for targeted audiences);
- Quality of the Peer Evaluation.

The Marking Scheme for Presentation Slides are provided on the next page for your reference.

This document is subject to change; however, announcement will be made be on the LEARN@PolyU homepage of this subject in case of any change has been made.
The group project will be based on the latest modified version of this document.

COMP5121 Data Mining & Data Warehousing Applications
Marking Scheme: Presentation Slides [Group Assignment]

Items	Marks	Remarks
Value(s) to your targeted audience	**20**	
• What are the objectives of the data analysis/mining to be performed in the perspective of your targeted audience? How does the objective relate to the problem that described in the case?	8	
• How are the results of the data analysis/mining meant to/being useful to your targeted audience?	6	
• How are the data analysis/mining and its results going to be used by your targeted audience?	6	
Details of the Design, Results and Discussion of the Data Analysis performed	**60**	
• Which steps of the KDD Process you have performed?	10	
• What are the design and implementation considerations that you have made for each step of the KDD Process performed, e.g. model adopted, parameters chosen? Describe how the chosen algorithm(s) is/are appropriate to obtain the expected result.	10	
• What are the results of each step of the KDD Process that you have performed? Describe the results of the data analysis/mining with respect to the objectives of the data analysis/mining.	15	
• How can the result of the data analysis/mining be used in the company? Based on the results obtained, are there any suggestion(s)/recommendation(s) can be made? (Please give logical and reasonable suggestion(s)/recommendation(s).	15	
• What do you expect in the company's business if the suggestion(s)/recommendation(s) are adopted by the senior management?	10	
Quality of the Presentation Slides	**20**	You will lose marks if you cramp too much text or information into one single slide.
• English fluency and presentation appropriateness	10	
• Professional slide formatting and layout	10	
Total:	100	
• Advanced/In-depth discussion,…, etc.	Bonus	

Scenario

The provided dataset is a sanitized version of operational data extracted from an operational database of your company. Assuming you are a group of data analysts and you are requested by the senior management of your company to perform Knowledge Discovery in Databases (KDD) process for your company.

You are known to be given limited time while you are not expected to provide a complete or thorough analysis. You are expected to provide some insights based on the provided dataset that could help the company decide on what should be done in the next step, e.g. a further, more in-depth analysis with other operational data that you require/recommend for such an analysis. You are expected to provide, if you could, useful interpretation(s) and suggestion(s) with some support from the provided Dataset.

Dataset/Data File

The dataset was recorded and extracted from an operational database of your company and provide to you in the form of a data file, `COMP5121_2016_3_Assignment-2_Part-III.csv`, in Comma-Separated Values (CSV) format.

The Dataset contains the sales data of the retail network of your company over the past 4 Months' time [with a total of 548 records]. The attributes/features of the dataset are as follows:

- AreaID

  A unique ID for different areas where there are retail stores located.
- AreaSize

  Size of the different areas where there are retail stores located.
- StoreID

  A unique ID for different retail stores.
- AgeOfStore

  Age of retail store in the unit of year.
- Promotion

  Promotion strategy used.
- Month

  A unique ID for the past 4 months.
- SalesInThousands

  Sales amount of a retail store in a particular month of time in the unit of thousand dollars ('000).

Note: The provided Dataset is the only set of information that is provided to you from your company at the moment and you cannot obtain any other further information and clarification from your company before your presentation or before the deadline of this group-based assessment.