EST 1892 LSBU

Coursework Specification

CW_Specification_CSI_6_DMA_23/24

Read this coursework specification carefully, it tells you how you are going to be assessed, how to submit your coursework on-time and how (and when) you'll receive your marks and feedback.

Module Code	CSI_6_DMA
Module Title	Data Mining and Big Data Analytics
Lecturer	Daqing Chen; George Bamfo
% of Module Mark	60%
Distributed	04/10/2023
Submission Method	Submit online via this Module's Moodle site
Submission Deadline	17:00, 11/12/2023
Release of Feedback & Marks	Feedback and provisional marks will be available in Grades on Moodle from 03/01/2024

Coursework Aim:

This individual coursework project is about using analytics to address real-world problems. The aim of this coursework is to evaluate your understanding of the basic theories, concepts, methodologies, and the various algorithms in data mining, and your skills of using Python for analytics tools.

Coursework Details:

Type:	Project report
Overall View:	This assignment is to be undertaken **individually**. You should plan your work thoroughly and have regular discussions with your module tutor to resolve any issues you may have during this project. The assignment involves analysing a real-world dataset and identifying its underlying structural patterns and models using appropriate data mining techniques and algorithms. These patterns and models are intended to be used to address certain business concerns. A dataset will be assigned to you by your tutor.

	You are expected to produce a written report for this data mining project.
Tasks:	You are required to undertake the following tasks in this project: 1. **<u>Business Understanding</u>** • Download the dataset assigned to you from the module Moodle site along with the data description file. • Read the data description file to learn the nature of the data, such as what is the data about, where it comes from, which certain business context it is associated with, etc. • Examine the dataset within its business context to identify meaningful problems that potentially can be addressed by using analytics. • Translate the business problems to appropriate data mining problems and tasks. • You may also refer to any articles published relevant to the dataset. 2. **<u>Data Understanding</u>** • Perform initial data exploration to get to know more about the dataset, such as the total number of instances in the dataset, the number of attributes (variables), the data type of each attribute, and the basic statistics of each attribute, including value range, average, standard deviation, skewness, kurtosis, and mode, etc. • Identify any data quality issues, including missing values, outliers, extreme values, incomparable value ranges of variables, and imbalanced classes, etc. • Determine if the dataset is appropriate to be used for addressing the business problems identified in Task 1. If not, re-do Task 1. 3. **<u>Data Preparation</u>** • Choose appropriate methods for data pre-processing, which includes dealing with missing values, tackling outliers, extreme values, and imbalanced classes, changing data types, reducing dimensionality, and conducting data transformation and normalisation, etc., wherever appropriate. • Identify correlations among certain variables.

	• Determine which and how each attribute should be used in your analysis. • Divide the whole dataset into several subsets to be used for training, test and validation in predictive modelling. 4. **Modelling** • Use the pre-processed dataset to perform the data mining tasks you have identified in Task 1. • Choose appropriate techniques and algorithms for your analysis: Choose either *k*-means clustering or association rule analysis for descriptive modelling, and choose either decision tree or regression for predictive modelling. • Determine appropriate settings of the algorithms to be applied, e.g., how many clusters to use in *k*-means clustering. • Re-do data preparation in Task 3 if needed. **5. Evaluation** • Provide an explicit and concise description and explanation of the descriptive and predictive models you have created. Examine and explain what patterns and insight have been identified. • Evaluate the performance of the predictive models in terms of evaluate the performance of the predictive models in terms of various measures applicable, such as accuracy, SSE (sum of squared errors), generalisation ability, simplicity and cost etc. • Discuss how the descriptive and predictive models created can be used to address the original business problems identified in Task 1. • Summarise your main findings from the project.
Word Count:	As a guide, aim for 2500-3000 words. The maximum word limit is 3000 words. If the total word limit is exceeded, it will affect the marks awarded to the project presentation. Footnotes will not count towards word count totals but must only be used for referencing, not for the provision of additional text. The bibliography will not count towards the word total.

Presentation:	• Report must contain Title page, Table of Contents, Abstract, Conclusion, and References. • Work must be referenced, and a bibliography provided. • Work must be submitted as a Word document (.doc/docx) or a PDF. • Course work must be submitted using Arial font size 11 (or larger if you need to), with a minimum of 1.5 line spacing. • Your student number must appear at the front of the coursework. Your name must **not** be on your coursework.
Referencing:	Harvard Referencing should be used, see your Library Subject Guide for guides and tips on referencing.
Regulations:	Make sure you understand the University Regulations on expected academic practice and academic misconduct. Note in particular: ▪ Your work must be your own. Markers will be attentive to both the plausibility of the sources provided as well as the consistency and approach to writing of the work. Simply, if you do the research and reading, and then write it up on your own, giving the reference to sources, you will approach the work in the appropriate way and will cause not give markers reason to question the authenticity of the work. ▪ All quotations must be credited and properly referenced. Paraphrasing is still regarded as plagiarism if you fail to acknowledge the source for the ideas being expressed. **TURNITIN:** When you upload your work to the Moodle site it will be checked by anti-plagiarism software.

Learning Outcomes

This coursework will partially assess the following learning outcomes for this module as indicated by *****.

Knowledge and Understanding
On successful completion of this module, you will be able to:

- Describe and explain the concepts of data mining including the techniques and algorithms for problem solving and creating competitive advantage. *****

Intellectual Skills
On successful completion of this module, you will be able to:

- Critically evaluate different types of data mining tasks in relation to various business and scientific problems, including descriptive modelling and predictive modelling, including cluster analysis, association analysis, and decision and regression for classification and prediction. *****

Practical Skills
On successful completion of this module, you will be able to:

- Transfer a business and/or scientific problem into an appropriate data mining problem. *****
- Creatively apply data mining tools and platforms such as Python package.*****

Transferable Skills
On successful completion of this module, you will be able to:

- Analyse and develop solutions for a wide range of business and scientific problems. *****

Assessment Criteria and Weighting

LSBU marking criteria have been developed to help tutors give you clear and helpful feedback on your work. They will be applied to your work to help you understand what you have accomplished, how any mark given was arrived at, and how you can improve your work in future.

	Criteria	Feedforward comments						
		100 - 80%	**79 - 70%**	**69 - 60%**	**59 - 50%**	**49 - 40%**	**39 - 30%**	**29 - 0%**
10%	**1. Business Understanding**	Exceptionally thorough and clear analysis of business concerns and associated data mining tasks.	Thorough and clear analysis of business concerns and associated data mining tasks.	Clear analysis of business concerns and associated data mining tasks to a certain depth.	Clear analysis of business concerns and associated data mining tasks. Probably lack some in-depth view.	Basic analysis of the key business concerns and associated data mining tasks.	Inadequate analysis of business concerns and associated data mining tasks. Lack clarity and relevance.	Little or no analysis of business concerns and associated data mining tasks.
10%	**2. Data Understanding**	Exceptionally excellent and creative initial data exploration with effective means. Thorough summary of the dataset. Excellent analysis of data quality issues and the role of each attribute. Excellent use of Python.	Excellent initial data exploration with effective means. Thorough summary of the dataset. Excellent analysis of data quality issues and the role of each attribute. Excellent use of Python.	Good initial data exploration performed with appropriate means. Clear summary of the dataset. Good analysis of data quality issues and the role of each attribute. Good and flexible use of Python.	Essential initial data exploration performed. Essential analysis of data quality issues and the role of each attribute. Good use of Python.	Limited simple initial data exploration. Probably lack some relevance and/or clarity. Limited use of Python.	Inadequate and/or inappropriate initial data exploration performed. Lack clarity and relevance. Poor use of Python.	Little or no initial data exploration performed. Little or no relevancy. No or inappropriate use of Python.
25%	**3. Data Pre-processing**	Exceptionally thorough and extensive consideration of data quality issues for pre-processing. Appropriate approaches adopted with exceptionally clear understanding. Excellent use of Python.	Thorough consideration of data quality issues for pre-processing. Appropriate approaches adopted with outstanding understanding. Excellent use of Python.	Good consideration of data quality issues for pre-processing. Appropriate approaches adopted with clear understanding and every aspect covered. Good and flexible use of Python.	Reasonable consideration of data quality issues for pre-processing. Appropriate approaches adopted with reasonable understanding and most of the main issues covered. Good use of Python.	Limited consideration of data quality issues for pre-processing. Some appropriate approaches adopted with limited understanding and limited coverage. Limited use of Python.	Inadequate and/or inappropriate view of data quality issues. Inappropriate approaches adopted. Poor use of Python.	Little or no data quality issues considered. Inappropriate approaches adopted. No or inappropriate use of Python.
15%	**4. Modelling**	Appropriate algorithms employed with exceptionally clear understanding. Modelling with excellent working knowledge of Python	Appropriate algorithms employed with clear understanding. Modelling with excellent working knowledge of SAS Enterprise Miner.	Appropriate algorithms employed with clear understanding. Good and flexible use of Python.	Appropriate algorithms employed with reasonable understanding. Good use of Python.	Some appropriate algorithms employed with limited understanding. Limited use of Python.	Inappropriate and/or inadequate algorithms employed. Poor use of Python.	Little or no algorithms employed. Little or no use of Python.
20%	**5. Model Evaluation**	Exceptionally thorough and clear model interpretation and comparison with regards to business concerns. Excellent and meaningful models/patterns created.	Thorough and clear model interpretation and comparison with regards to business concerns. Excellent meaningful models/patterns created.	Clear model interpretation and comparison with regards to business concerns. Significantly meaningful models/patterns created.	Basic model interpretation and comparison with regards to business concerns. Reasonable models/patterns created.	Weak model interpretation and comparison with regards to business concerns. Very limited meaningfulness. Probably lack some clarity.	Poor model interpretation and comparison with regards to business concerns. No or little meaningful models/patterns provided.	Little or no model interpretation and comparison with regards to business concerns.
20%	**6. Report**	Exceptionally clear and concise summary of project findings. May raise questions for future research. Exceptional outstanding presentation. Clear structure and layout.	Very clear and concise summary of project findings. May raise questions for future research. Outstanding presentation. Clear structure and layout.	Clear and concise summary of project findings. Excellent presentation. Clear structure and layout.	Clear review and summary of project findings. Good presentation with proper structure and layout.	Adequate review of project findings. Probably lack of some clarity. Acceptable presentation.	Inadequate review of project findings. Lack of clarity and accuracy. Poor presentation.	Little or no review of project findings. Significantly Lack of clarity and accuracy. Very poor presentation.

How to get help

We will discuss this Coursework Specification in class. However, if you have related questions, please contact me [name and email] as soon as possible.

Resources

List resources such as background reading, templates, samples, tools, videos, links, etc

Quality assurance of coursework specifications

Coursework specifications within CSI division go through internal (for new modules with 100% coursework also through external) moderation. This is to ensure high quality, consistency and appropriateness of the coursework as well as to share best practice within the CSI division.