EST 1892 LSBU

Coursework Specification

CW_Specification_CSI_7_DMA_2023-24

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_7_DMA
Module Title	Data Mining and Analysis
Lecturer	Daqing Chen, Kamal Thapa, George Bamfo, Godwin Idoje
% of Module Mark	60%
Distributed	24/01/2024
Submission Method	Online submission via the module's Moodle site on the VLE
Submission Deadline	Stage report: 17:00, Friday 16/02/24 Final report: 17:00, Friday 26/04/24
Release of Feedback & Marks	Feedback and provisional marks will be available in Gradebook on the VLE from 13/05/2023

Coursework Aim:

This individual project assessment requires you to analyse a real-world dataset and provide meaningful insights from your analysis aiming to address certain business concerns and problems.

Coursework Details:

Type:	Report
Overview	The objective of this individual assignment is to evaluate your understanding of the basic theory, concepts, and various methods and algorithms in data mining, and assess your skills of applying appropriate Python packages, such as NumPy, Pandas, Matplotlib, and Scikit-learn, etc., to carry out a data mining project. The dataset for this coursework is related to London Fire Brigade (LFB) incidents reported from 2019 to 2022. Your role in this project is two-fold: acting as a business client and as a data analyst. As a business client, you are expected to raise meaningful business concerns/problems in relation to the data given. And as a data analyst, you are required to follow a proper data mining methodology and apply various techniques covered in lectures and tutorials to analyse your data to address the business concerns and problems having been raised.
Tasks	You are required to undertake the following tasks: 1. Problem Identification 1.1. Read the data description file (metadata) to learn the context and the basic characteristics of the dataset assigned to you. 1.2. Identify 3 – 5 meaningful business problems of interest with regard to the data for analysis. 1.3. Identify what data mining tasks need to be performed in order to address the business problems raised. 2. Data Understanding 2.1. Conduct initial data exploration with essential EDA (exploratory data analysis, graphs/statistics) to learn more about the given dataset, including the total number of variables (dimensions, attributes, fields), the data type of each variable, the value range/mode, skewness, and kurtosis of each variable, and the total number of instances (records, tuples), etc. 2.2. Identify any data quality issues in the given dataset, including missing values, outliers, extreme values, imbalanced proportions (classes, categories) of categorical variables, and incomparable value ranges of numeric variables, 2.3. Consider if the given dataset is appropriate and sufficient to be used for addressing the business problems identified in Task 1 and re-do Task 1, if not.

	Note: Stage report should include Tasks 1 and 2. 3. Data Preparation 3.1. Determine which variables to be used in which analysis. Also refer to 1.2. and 1.3. in Task 1. 3.2. Get your data ready for analysis. Choose appropriate methods for data pre-processing, i.e., deal with incorrect data types, drop/reject irrelevant variables, tackle missing values, outliers, imbalanced classes, and duplicates; change data type, dimensionality reduction, feature extraction, data transformation, normalisation, and data partition, etc. wherever appropriate. Also refer to 1.1 in Task 1, and 2.2 in Task 2. 4. Model Construction 4.1. With the pre-processed dataset undertake the data mining tasks you have identified in 1.2. You are required to apply **two different algorithms for both predictive and descriptive modelling**. For descriptive modelling, you may choose to use the *k*-means clustering and various EDA methods, e. g., histograms, bar charts, and Person's correlation coefficient, etc. For predictive modelling, you may use decision trees and artificial neural networks, or decision trees and *k*-nearest-neighbour, etc. 4.2. In order to build the most appropriate and accurate models to identify meaningful hidden patterns, different settings for the relevant model parameters should be considered for each of the selected algorithms and methods. 5. Model Interpretation and Evaluation 5.1. Interpret the descriptive models created, such as clusters created using *k*-means algorithms, correlation among variables, and various relevant plots created. 5.2. Compare the performances of different predictive models in terms of accuracy, error rate, generalisation capability (over-fitting), simplicity and cost, etc., where appropriate. 5.3. Discuss the meaningfulness and usefulness of the models built and the patterns revealed, and **how the models and the patterns can be used to address the original business concerns**. This includes both descriptive and predictive models. 6. A summary of the main findings of the project and your suggestion(s) to LFB based on your analysis.
Word Count:	As a guide, aim for 3000 - 3200 words: Stage report 700 -

	800 words; Final report 2300 - 2400, excluding Title page, Table of Contents, tables and graphs, footnotes, bibliography, and scripts. The maximum word limit is 3200 words. You may get a reduction in mark for not meeting the word count limit.
Presentation:	• 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 and business analytics.
- Critically review and appreciate the role of data mining in business analytics. *
- Critically explain how and why data mining and business analytics can be used to create competitive advantage for businesses and enterprises. *

- Critically analyse when, why, and how data mining should be considered a possible problem-solving strategy from a business perspective. *
- Gain sufficient working knowledge of using Python packages and libraries, such as Numpy, Pandas, Matplotlib, and Sklearn, etc., for performing data exploration, detecting and data quality issues, modelling, model interpretation and comparison, and reporting with real-world case studies. *

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

- Identify different types of data mining tasks in relation to various business concerns, including classification, prediction, cluster analysis and segmentation, and association analysis and market basket analysis. *
- Critically review and appreciate the strengths and weaknesses of different data mining techniques, models, and tools. *****

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

- Select and apply appropriate data mining techniques for a given real-world problem. *
- Evaluate various models built from a data mining process. *
- Undertake a data mining project with clear business focus, in particular, in relation to CRM analysis, RFM modelling, and credit risk scoring. *****

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

- Demonstrate analytical skills. *
- Demonstrate project management skills. *
- Teamwork skills. *****

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%
15%	**1. Business Understanding and Data Understanding**	Exceptionally clear and concise analysis of business concerns and relevant data mining tasks. Excellent and creative initial data exploration with effective means. Thorough and extensive consideration of data quality issues.	Thorough and clear analysis of business concerns and relevant data mining tasks. Excellent initial data exploration with effective means. Thorough consideration of data quality issues.	Clear analysis of business concerns and relevant data mining tasks to a certain depth. Sensible initial data exploration performed with appropriate means. Good consideration of data quality issues.	Clear analysis of business concerns and relevant data mining tasks. Probably lack some in-depth view. Essential initial data exploration performed. Reasonable consideration of data quality issues.	Adequate analysis of the key business concerns and data mining tasks. Limited initial data exploration. Probably lack some relevance. Inappropriate means. Limited consideration of data quality issues.	Inadequate analysis of business concerns and data mining tasks. Only simple initial data exploration performed. Lack clarity and relevance. Inadequate view of data quality issues.	Little or no analysis of business concerns and data mining tasks. Little or no initial data exploration performed. Little or no relevancy. Little or no data quality issues considered.
25%	**2. Data Pre-processing**	Thorough and appropriate approaches adopted for data pre-processing with exceptionally clear understanding. Excellent use of the relevant Python packages.	Appropriate approaches adopted for data pre-processing with outstanding understanding. Excellent use of the relevant Python packages.	Appropriate approaches adopted for data pre-processing with clear understanding and every aspect covered. Good and flexible use of the relevant Python packages.	Appropriate approaches adopted for data pre-processing with reasonable understanding and most of the main issues covered. Good use of the relevant Python packages.	Some appropriate approaches adopted for data pre-processing with limited understanding and limited coverage. Limited use of the relevant Python packages.	Inappropriate approaches adopted for data pre-processing. Poor use of the relevant Python packages.	Inappropriate approaches adopted for data pre-processing. No or inappropriate use of the relevant Python packages.
15%	**3. Model Construction**	Appropriate algorithms employed with Exceptionally clear outstanding understanding. Modelling with excellent working knowledge of the relevant Python packages.	Appropriate algorithms employed with outstanding understanding. Modelling with excellent working knowledge of the relevant Python packages.	Appropriate algorithms employed with clear understanding. Good and flexible use of the relevant Python packages.	Appropriate algorithms employed with reasonable understanding. Good use of the relevant Python packages.	Some appropriate algorithms employed with limited understanding. Limited use of the relevant Python packages.	Inappropriate algorithms employed. Poor use of the relevant Python packages.	Little or no algorithms employed. Little or no use of the relevant Python packages.
25%	**4. 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%	**5. 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

If you have related questions, please contact Daqing Chen, email: chend@lsbu.ac.uk, as soon as possible.

Resources

All the module’s lectures, tutorial handouts, and the references recommended in the module guide.

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.