THE UNIVERSITY OF WESTERN AUSTRALIA

ECM

**SEMESTER 2, 2017 SUPPLEMENTARY AND DEFERRED EXAMINATIONS**

**CITS4009**
**Introduction to Data Science**

FAMILY NAME: _____ GIVEN NAMES: _____

STUDENT ID: | | | | | | | | | SIGNATURE: _____

This Paper Contains: **5** pages **(including title page)**
Time allowed: **2:00 hours**

INSTRUCTIONS:

UWA approved calculators are allowed.
Students should answer all questions.

**PLEASE NOTE**

Examination candidates may only bring authorised materials into the examination room. If a supervisor finds, during the examination, that you have unauthorised material, in whatever form, in the vicinity of your desk or on your person, whether in the examination room or the toilets or en route to/from the toilets, the matter will be reported to the head of school and disciplinary action will normally be taken against you. This action may result in your being deprived of any credit for this examination or even, in some cases, for the whole unit. This will apply regardless of whether the material has been used at the time it is found.

Therefore, any candidate who has brought any unauthorised material whatsoever into the examination room should declare it to the supervisor immediately. Candidates who are uncertain whether any material is authorised should ask the supervisor for clarification.

Supervisors Only Student left at:

This page has been left intentionally blank

**Q1.**

**(a)**

Explain the different roles in a data science project. Why is it important to set precise quantitative goals for a data science project?

(6)

**(b)**

Explain clearly the different stages of a data science project. Which stage do you think is the most important? How would you approach this stage? Explain through an example.

(6)

**Q2.**

**(a)**

Explain what are *missing values* and *outliers* in data. Explain how you will deal with missing values and outliers in a data science project.

(6)

**(b)**

What are the benefits of checking distributions for a single variable in a large dataset? What are the key aspects that you think are important to look for in such a visualization? Explain clearly.

(6)

**Q3.**

**(a)**

Explain clearly the different parts of the model construction process. What are test and training data? How would you generate test and training data for a data science project?

(6)

**(b)**

You have been recruited by an e-commerce company as a data scientist. The company sells many different products including books, electronic products, kitchen accessories, home improvement products etc. Each customer needs to create an account at the web protal of the company and supply some personal information like age, gender and location.

Your job is to recommend products to users while they are browsing the catalogue, so that users are attracted to purchase the recommended products. The company stores historical data for the users, e.g., which products they have browsed, and which ones they have purchased. Explain clearly a strategy that you will use for recommending appropriate products to the users to improve the profit of the company.

(6)

**Q4.**

**(a)**

Explain the difference between *supervised* and *unsupervised* learning. Give examples of business data analysis tasks for each of these learning methods.

(6)

**(b)**

Explain the meaning of the two terms *specificity* and *recall* in relation to evaluating classification models. Write the mathematical expressions for these two measures. Explain the terms in your mathematical expressions.

(6)

**Q5.**

**(a)** What are the key points that you should include in a presentation to the sponsor of a data science project? Explain clearly why these points are important. What are the key points that you should include in a presentation to the end users of a data science project?

(6)

**(b)**

Explain when *regression* methods are useful for data modeling. What is the difference between *linear* and *logistic* regression? Explain with examples.

(6)

———————————-END OF PAPER———————————