Faculty of Computing and Information Technology/Department of Computer Science and Mathematics

BACS3013 Data Science

Tutorial 1 (Introduction to Data Science and Big Data Analytics)

Q1. Data are everywhere. A main obstacle to fully harnessing the power of big data using analytics is the lack of skilled resources and "data scientist" talent required to exploit big data.

(a) Give **five (5)** examples on how data scientists may turn data into insight. (15 marks)

An online social media asks its users to list their hometown and current location, to make it easier for their friends to find connect with them. But it may also analyse these locations to identify global migration patterns. (3m)

A (large) retailer tracks customers' purchases and interactions, both online and instore. It may use the data to predictively which of its customers are pregnant, to better market baby-related purchases to the customers. (3m)

A dating site asks its members to answer thousands of questions in order to find the most appropriate matches for them. (3m)

A financial institution requests its customers to complete a behavioural scoring model to predict the credit risk. (3m)

A supermarket analyses its customers market baskets to decide on product bundling, next best offer, and improve shelf organization. (3m)

(b) Basic unit of analysis plays important role in data analytics. In the context of data analytics, describe **two (2)** examples on how individuals can be the unit of analysis.

(6 marks)

In descriptive studies, the data scientists make observation describing the characteristics of a large number of individual people in using online social media, such as their genders, ages, attitudes etc. (3m)

A research project examines whether secondary school graduates in rich families are more likely to attend college/University than those in poor families. The individual graduate is the unit of analysis (not rich or poor families) (3m)

BACS3013 Data Science

Q2. List out **five (5)** job profiles that need to collaborate together in data analytics. For each job profile, briefly describe the roles that they are playing. (15 marks)

A database or data warehouse administrator (DBA) is aware of all the data available within the firm, the storage details, and the data definitions. (3m)

A business expert (eg: credit portfolio manager) has extensive business experience and common sense to steer the analytical modeling exercise and interpret its key findings. (3m)

A legal expert has good knowledge about what data can be used when, and what regulation applies in what location. For instance, in credit risk modeling, one can typically not discriminate good and bad customers based upon gender, race, or religion. (3m)

A data scientist is responsible for doing the actual analytics. He should possess a thorough understanding of all techniques involved and know how to implement them using the appropriate software. (3m)

A software tool vendor provides tool to automate specific steps of the analytical modelling process (e: data preprocessing). Some vendors might provide analytics-based solution for specific application areas, such as risk management and marketing analytics. (3m)

Q3. A good analytical model should satisfy several requirements. describe the importance of each requirement or factor listed below.

(a) Business relevance (2 marks)

The analytical model should actually solve the business problem for which it was developed. It makes no sense to have a working analytical model that got sidetracked from the original problem statement. (2m)

(b) Statistical performance (2 marks)

The model should have statistical significance and predictive power. For example, in a classification setting, the model should have good discrimination power. (2m)

(c) Interpretability (2 marks)

The patterns that the analytical model captures must be interpretable. For example, in credit risk modeling, interpretable models are needed to get good insight into the underlying data patterns. (2m)

(d) justifiability (2 marks)

Justifiability refers to the degree to which model corresponds to prior business knowledge and intuition. For example, a model stating that a higher debt ratio results in more creditworthy may be interpretable (due to the limited/biased data samples), but is not justifiable because it contradicts basic financial intuition. (2m)

(e) operational efficiency (2 marks)

It refers to the efforts needed to collect the data, preprocess it, evaluate the model a nd feed its outputs to the business application. For instance, in a real-time crowd counting application and analysis, operational efficiency may be a crucial characteristic. (2m)

(f) economic cost (2 marks)

The costs to gather and preprocess the data, the costs to analye the data, and the costs to put the resulting analytical models into production have to be taken into consideration. It is important to do a thorough cost-benefit analysis at the start of the project. (2m)

(g) compliance with the regulation and legislation (2 marks)

Analytical models should comply with both local and international regulation and legislation. For instance, under the Personal Data Protection Act 2010, data users and data processors are not allowed to process the sensitive personal data except for the purposes specified in the Art and must be with explicit consent of the data subject [1].

Another example is , in a credit risk setting, the Basel II and Basel III Capital Accords have been introduced to appropriately identify the types of data that can (or cannot) be used to build credit risk models.

[1] Department of Personal data protection. 2017. *FREQUENTLY ASKED QUESTION*. [ONLINE] Available at: http://www.pdp.gov.my/index.php/en/soalan-lazim**.**

Q4. List out four (4) types of analytics. (4 marks)

Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics. (4m)

Q5. For each type of analytics shown below, A good analytical model should satisfy several

requirements. describe the importance of each requirement or factor listed below.

(a) Descriptive Analytics (2 marks)

Descriptive Analytics is a form of advanced analytics which examines data or content to answer the question "What happened?" (2)

(b) Diagnostic Analytics (2 marks)

Descriptive Analytics is a form of advanced analytics which examines data or content to answer the question "Why did it happen?" (2)

(c) Predictive Analytics (2 marks)

Descriptive Analytics is a form of advanced analytics which examines data or content to answer the question "What could happen in the future?" (2)

(d) Prescriptive Analytics (2 marks)

Descriptive Analytics is a form of advanced analytics which examines data or content to answer the question "How should we respond to those potential future events?" (2)

Q6. A toy company experienced a huge lose last year. A data scientist is employed to help the company to analyze the possible reasons and propose a realistic profit plan. Justify if the following types of analytics is helpful in solving the problem.

(a) Descriptive Analytics (3 marks)

Descriptive analytics may report historical insights regarding the company's production, financials, operations, sales, finance, inventory and customers. It is helpful in understanding at an aggregate level what is going on in that company. (3)

(e) Predictive Analytics (2 marks)

Predictive Analytics may combine the historical data found in database/data warehouse to identify patterns in the data and apply mathematical models and algorithms to capture relationships between various datasets. Hence, predictive analytics can be used throughout the company, from forecasting customer behaviors to identifying trends in sale activities. (3m)

(f) Prescriptive Analytics (2 marks)

Prescriptive Analytics goes beyond descriptive and predictive analytics by recommending one or more possible courses of action. It is helpful in optimizing production, scheduling and inventory in the supply chain; and optimizing the customer experience. (3m)

BACS3013 Data Science