Swinburne University of Technology Sarawak Campus
Faculty of Engineering, Computing and Science

COS10022 Introduction to Data Science
Assignment 1 - Semester 2, 2020

Assessment Title:
Framing a business problem into a data analytic problem

Assessment Weighting:
20%

Due Date:
Monday, 12th October 2020 at 11.59 pm (Malaysian Local Time)

Assessable Items:

1. One (1) written report of **not more than** 10-page long with the Assignment Cover Sheet.
2. One (1) Excel file containing **not more than** 5 spreadsheets.

Submission Requirements:

1. ADD headers to all pages of the report, which should include your group number, unit code and assessment title.
2. The length of the report should NOT be more than 10 pages, including the reference page (single line space; 11pt font; Arial).
3. The length of the Excel file should NOT contain more than 5 spreadsheets.
4. Title Page and Table of Content is NOT required for the report.
5. CITE all references using Harvard Referencing Style.
6. CREATE a table showing the distribution of work done by each team member on the last page of the report.
7. NAME your report as 'Group Number_A1_Report.pdf' (e.g. Group 1_A1_Report.pdf).
8. NAME your Excel file as 'Group Number_A1_Analysis.xlsx'
   (e.g. Group 1_A1_Analysis.pdf).
9. SUBMIT the three assessable items via the Canvas unit site: (a) Combine the two assessable items into a .zip folder, and (b) Upload the .zip folder to Canvas

Purpose of Assignment:

This is a group assignment. This assignment is to be completed in a group of 3 to 4 students. Unless students make an explicit request to differ, the group mark will be distributed equally among all group members (refer to the Unit Outline for the late submission penalty and group work policy). This assignment aims at evaluating students' achievement of the following unit learning outcomes:

1. **Appreciate (and discuss) the roles of data science and Big Data analytics in business and organisational contexts.**
2. **Appreciate (and explain) the key concepts, techniques and tools for discovering, analysing, visualising and presenting data.**
3. **Describe the processes within the Data Analytics Lifecycle**.

Key Lessons:

To offer real commercial values, the practices of data science should address a real and specific business problem. This requires substantial understanding of the nature of the business problem at hand, as well as developing the ability to frame business problems into analytics problems.

Business Case:

"Optimizing Product Placement in Retail"

BigMart Sales Dataset is one of the popular datasets available on Kaggle website. Assume that a group of data scientists at BigMart have collected 8,523 sales data for 1,559 products across 10 stores in different cities. Certain attributes of each of the products and stores have also been defined. The aim is to build a predictive model and identify the sales of each product at a particular store. Using this model, BigMart will try to understand the properties of products and/or stores which play a significant role in increasing sales.

Assignment Tasks:

Your task is to produce a data science proposal that discusses the possibility of automatically predicting the "sales of a product". The length of the report should not be more than 10 pages, including the title page and the reference page (single line space; 11pt font; Arial). Table of Content is not required. There are 100 marks in this assignment.

Your proposal must address the following tasks:

1. [30 marks] The file ***BigMart.csv*** contains a dataset of 1,559 products across 10 stores in different cities. The description of the dataset is provided in the **Data Dictionary** section. Use the following steps to formulate appropriate hypotheses about the expected outcomes from the analysis of the dataset:

   (a) Identify the dependent variable and the independent variables, and determine their data types (nominal, ordinal, continuous or discrete);

   (b) Discuss the relevance of each independent variable for the prediction of the dependent variable; and

   (c) Develop five (5) store level or product level hypotheses based on the discussion in Step 1(b). For instance, "Stores located in city areas have higher sales because of higher demand for household products".

2. [40 marks] To build a prediction model requires that we have a training dataset and a test dataset. The training dataset provides a predictive model with the actual outcomes to learn from while the test dataset hides the actual outcomes from the model and serves as the basis for measuring the model's prediction accuracy. Perform the following steps to construct a training dataset and a test dataset for the sales prediction problem:

   (a) Make a copy of the ***sample*** worksheet in the ***BigMart.csv*** file and rename the new worksheet as ***pre-processing.***

   (b) Impute variables with missing or invalid data on the ***pre-processing*** worksheet. Explain the imputation strategies and steps involved.

   (c) Create meaningful categorical data out of the existing numerical data on the***pre-processing*** worksheet. Explain the reasons behind and the pre-processing steps involved.

   (d) Make a copy of the ***pre-processing*** worksheet and rename the new worksheet as ***train***.

   (e) Produce a test dataset from the training dataset and rename the new worksheet as ***test***. Explain the strategies and steps involved in this step.

   (f) Rename the workbook as ***BigMart_Your Group Number.csv*** and save it in a new folder. Your workbook should contain a total of four worksheets ('Sample', 'pre-processing', 'train' and 'test').

3. [10 marks] Discuss the type of output variable (i.e. the dependent variable) to be predicted. For instance: should it be a categorical or a continuous variable? Specifically, you are required to discuss the pros and cons of treating the output variable as categorical or continuous.

4. [15 marks] Justify for several business values to be gained from the ability to automatically predict the expected sales of a product.

5. [5 marks] Present all your answers in the form of a high-quality written report.

6. Create a table showing the distribution of work done by each team member on the assignment.

PLEASE READ ME

Do I need to do the actual prediction of the sales of products?

No. You DO NOT need to create any data science model to perform any actual prediction. The proposal only describes your idea.

Do I need to employ any programming in this assignment?

No. Coding skills is irrelevant to this assignment and shall not contribute additional marks.

The *BigMart.csv* file is created based on the 'BigMart Sales' dataset available at: http://www.kaggle.com/. There are already abundant works dedicated to studying the problem of predicting sales of products using machine learning and artificial intelligence methods. Similar works can be found online. You are encouraged to explore some of the existing literature and, where applicable, adapt their ideas into your work. When you do so, please include all the necessary in-text citations and the end-of-report reference list. The Harvard Referencing format must be used when citing and referencing external information resources: https://www.swinburne.edu.my/library/referencing/harvard-style-guide.php.

Data Dictionary:

Column	Attribute Name	Definition	Example	% Null Ratios
A	Item_Identifier	It is a unique product ID assigned to every distinct item. It consists of an alphanumeric string of length 5.	FDN15	0
B	Item_Weight	This field includes the weight of the product.	17.5	17.17
C	Item_Fat_Content	This attribute is categorical and describes whether the product is low fat or not. There are 2 categories of this attribute: ['Low Fat', 'Regular']. However, it is important to note that 'Low Fat' has also been written as 'low fat' and 'LF' in dataset, whereas, 'Regular' has been referred as 'reg' as well.	Low Fat	0
D	Item_Visibility	This field mentions the percentage of total display area of all products in a store allocated to the particular product.	0.01676	0
E	Item_Type	This is a categorical attribute and describes the food category to which the item belongs. There are 16 different categories listed as follows: ['Dairy', 'Soft Drinks', 'Meat', 'Fruits and Vegetables', 'Household', 'Baking Goods', 'Snack Foods', 'Frozen Foods', 'Breakfast', 'Health and Hygiene', 'Hard Drinks', 'Canned', 'Breads', 'Starchy Foods', 'Others', 'Seafood'].	Meat	0
F	Item_MRP	This is the Maximum Retail Price (list price) of the product.	141.618	0
G	Outlet_Identifier	It is a unique store ID assigned. It consists of an alphanumeric string of length 6.	OUT049	0
H	Outlet_Establishment_Year	This attribute mentions the year in which store was established.	1998	0
I	Outlet_Size	The attribute tells the size of the store in terms of ground area covered. It is a categorical value and described in 3 categories: ['High', 'Medium', 'Small'].	Medium	28.27642849
J	Outlet_Location_Type	This field has categorical data and tells about the size of the city in which the store is located through 3 categories: ['Tier 1', 'Tier 2', 'Tier 3'].	Tier 3	0
K	Outlet_Type	This field contains categorical value and tells whether the outlet is just a grocery store or some sort of supermarket. Following are the 4 categories in which the data is divided: ['Supermarket Type1', 'Supermarket Type2', 'Grocery Store', 'Supermarket Type3'].	Supermarket Type2	0
L	Item_Outlet_Sales	This is the outcome variable to be predicted. It contains the sales of the product in the particular store.	2097.27	0