Swinburne University of Technology

COS10022 Introduction to Data Science

Assignment 1

Assessment Title: Framing a Business Problem into a Data Analytic Problem

Assessment Weighting: 20%

Due Date: Refer to Canvas

Assessable Items:

1. One (1) piece of written report of **not more than** 10-page long with the Assignment Cover Sheet (digitally signed by all group members).
2. Three (3) files containing the cleaned, the training, and the test dataset (in csv format)
3. One (1) file containing the process log of what you did to your input data, the sequence you did in the data preparation process. (in docx format). If the preparation process is executed in KNIME platform, you can include the screenshot of the workflow in the log.

You must include a digitally signed Assignment Cover Sheet with your submission.

Purpose of Assignment

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**.

This is a **group** assignment. This assignment is to be completed in a group of three (**3)** 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.

"Optimizing Product Placement in Retail"

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.

Introduction

BigMart Sales Dataset is one of the popular dataset collected on Kaggle website. The data scientists at BigMart have collected 8,523 sales data for 1,559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store. Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.

Please note that the data may have missing values as some stores might not report all the data due to technical glitches. Hence, it will be required to treat them accordingly.

Assignment Goal

One of the datasets from BigMart (the training dataset) is given to you. It contains 8,523 sales records (rows) with 12 columns (attributes) collected from different outlet stores with the item lists. The aim is to build a predictive model and identify the sales of each product at a particular store. The predictive model will help BigMart to understand the properties of products and/or stores which play a significant role in increasing sales.

Assignment Task

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. The file *Assignment1_BigMart_Data.csv* (available on Canvas) a dataset of 1,559 products across 10 stores in different cities. The description of the dataset is provided in the appendix. Use the following steps to formulate appropriate hypotheses about the expected outcomes from the analysis of the dataset:

    a. Clean the dataset to remove/patch the missing records/values and record all changes.
    b. Identify the dependent variable and the independent variables, and determine their data types (nominal, ordinal, continuous or discrete);
    c. Discuss the relevance of each independent variable for the prediction of the dependent variable; and
    d. Develop five (5) store level or product level hypotheses based on the discussion in Step 1(c). For instance, "Stores located in city areas have higher sales because of higher demand for household products".

    [30 marks]

2. 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 *Assignment1_BigMart_Data.csv* file and rename the new worksheet as ***pre-processing.csv***.
   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. Split 70% of the data and save it as ***train.csv***.
   e. Collect the remaining 30% data, remove the prediction target attribute and save it as ***test.csv***. Explain the strategies and steps involved in this step.

   Rename *pre-processing*.csv, train.csv, and test.csv as **BigMart_pre_oo.csv, BigMart_tra_oo.csv, and BigMart_tes_oo.csv**, respectively and save them, where **oo** should be replaced by a 2-digit group number.
   [40 marks]

3. 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.
   [10 marks]

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

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

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

The Assignment1_BigMart_Data.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: http://www.swinburne.edu.au/library/referencing/harvard-style-guide/

PLEASE READ ME

Why this assignment?
Completing this assignment helps you to develop skills in:

- Framing a business problem into a data science problem (Task 1 and Task 2)
- Data collection (Task 3)
- Justifying the business value of a proposed data science solution (Task 4)

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.

Submission Requirement

To fulfil the requirement of this assignment, the following items must be submitted and packaged into a single .ZIP file named ***COS10022_("Your Group Number")*** and submitted:

- One (1) piece of written report of **not more than** 10-page long with the Assignment Cover Sheet (digitally signed by all group members).
- Three (3) files containing the cleaned, the training, and the test dataset (in csv format)
- One (1) file containing the process log of what you did to your input data, the sequence you did in the data preparation process. (in docx format). If the preparation process is executed in KNIME platform, you can include the screenshot of the workflow in the log.

Failure to adhere to the submission requirements will immediately result in "N" grade for this assignment.

Rubric: Subtask 1

Marks	Data Structure [5 marks]	Data Exploration [20 marks]	Hypothesis Generation [5 marks]
24.0 – 30.0	Correctly identify the dependent variable [1 mark] and the independent variables from the *sample* worksheet [1 mark]. Correctly determine the data types of all variables in the *sample* worksheet [0.25 each x 12 = 3 marks].	Evidence of higher cognitive activities such as contrasting and arguing based on the research materials. [10 marks] All explanations of the variables' relevance are correct/relevant, and all are supported by citations to external sources. [10 marks]	Each hypothesis carries 1 mark. The correctness of a hypothesis is judged based on the explanation of the relevance of an independent variable for the prediction of the dependent variable. [5 hypotheses = 5 marks].
21.0 - 23.9		All explanations of the variables' relevance are correct/relevant.	
18.0 - 20.9		Some explanations of the variables' relevance are incorrect/irrelevant.	
15.0 – 17.9		All variables include explanations of their relevance.	
0.0 – 14.9		25% for incomplete explanation of relevance 0% for no explanation of relevance at all	

Rubric: Subtask 2

Marks	Sample Dataset (40 marks)
32.0 – 40.0	There are detailed explanations of the imputation methods used to handle missing or invalid data (Question 2b). [10 marks] There are detailed explanations of the transformation of numerical data to categorical data (Question 2c). [10 marks] There are detailed descriptions of how the test dataset was derived randomly using Excel formulas from the training dataset (Question 2e). [10 marks] All explanations are correct, supported by relevant external sources which are properly cited and referenced. [10 marks]
28.0 – 31.9	There are descriptions of the data pre-processing steps involved in producing the training dataset and the test dataset (Questions 2b, 2c, 2e), and training set has imputed values in place of missing/invalid data and the test dataset is randomly selected from the training dataset. All explanations are correct, but part of the explanations is not supported by relevant external sources which are properly cited and referenced.
24.0 – 27.9	Some descriptions of the pre-processing steps involved are incomplete (Questions 2b, 2c, 2e), but training set has imputed values in place of missing/invalid data and the test dataset is randomly selected from the training dataset. Some explanations are incorrect/irrelevant and are not supported by relevant external sources which are properly cited and referenced.
20.0 – 23.9	Poor descriptions of the pre-processing steps involved (Questions 2b, 2c, 2e), and the training dataset has incomplete or invalid values and the test dataset is manually generated. Some explanations are not provided and are not supported by relevant external sources which are properly cited and referenced.
0.0 – 19.9	25%: incomplete datasets that does not satisfy the minimum requirement of the assignment. 0%: no dataset; inaccessible datasets.

Rubric: Subtask 3

Marks	Output Variable (10 marks)
8.0 – 10.0	Discuss relevant data science issues beyond what the direct requirements of the question. All are highly appropriate and accurate.
7.0 – 7.9	Demonstrate *attempts* to discuss relevant data science issues beyond what the direct requirements of the question. Some discussions may be inappropriate.
6.0 – 6.9	The answer considers more than one possible types of output variables, supported with relevant rationale. The answer to the question considers the pros and cons of predicting a single winner.
5.0 – 5.9	The answer only considers one type of output variable, supported with relevant rationale.
0.0 – 4.9	25%: students attempt to address the question but most of the answer is irrelevant. 0%: no answer; or completely irrelevant answer.

Rubric: Subtask 4

Marks	Business Values (15 marks)
12.0 – 15.0	Each argument is supported by references to examples/real life cases.
10.5 – 11.9	Identify 3 distinct business values.
9.0 – 10.4	Business values are clearly identified. Connections between the benefits of a predictive modelling and each of the business value are clearly argued and elaborated.
7.5 – 8.9	Business values are clearly identified. Most of the discussions are based only on the students' personal opinions.
0.0 – 7.4	25%: most of the discussions are irrelevant. 0%: no discussions at all.

Rubric: Subtask 5

Marks	Quality of the Written Report (5 marks)
4.0 – 5.0	Ideas are coherently structured and presented. Effective use of tables and/or figures.
3.5 – 3.9	Consistent uses of in-text citations. References are present in a consistent **Swinburne's Harvard referencing format**.
3.0 – 3.4	The length of the report does not exceed 10 pages. The report has a clear structure. References are present but use inconsistent reference style.
2.5 – 2.9	The report uses consistent formatting. (e.g. consistent font type, font size, etc.) The report is written in English and readable. Some grammatical mistakes may be present.
0.0 – 2.4	25%: The report does not reflect a professional standard. 0%: Most of the report is unreadable.

Appendix

Data Dictionary

Column Position	Attribute Name	Definition	Example	% Null Ratios
1	Item_Identifier	It is a unique product ID assigned to every distinct item. It consists of an alphanumeric string of length 5.	FDN15	0
2	Item_Weight	This field includes the weight of the product.	17.5	17.17
3	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
4	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
5	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
6	Item_MRP	This is the Maximum Retail Price (list price)	141.618	0

		of the product.		
7	Outlet_Identifier	It is a unique store ID assigned. It consists of an alphanumeric string of length 6.	OUT049	0
8	Outlet_Establishment_Year	This attribute mentions the year in which store was established.	1998	0
9	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
10	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
11	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
12	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