Swinburne University of Technology Hawthorn Campus
Dept. of Computer Science and Software Engineering

COS10022 Introduction to Data Science

Assignment 1 - *Semester 1, 2021*

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

Assessment Weighting: 20%

Due Date: Tuesday, 6th April 2021 at 11.59 pm (AEDT)

Assessable Item:

- One (1) piece of a written report of not more than 20-page long with the Assignment Cover Sheet, which should be digitally signed by all group members.

The submitted report should have a section describing the number of records, the attributes, and all characteristics of the cleaned, the training, and the test dataset, respectively. Screenshots can be used for helping you describe your processed result. Your report should also reveal what you did to your input data and the sequence you did in the data preparation process. If the preparation process is executed in the KNIME platform, you can include the screenshot of the workflow.
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 five (**5**) students. 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. You can ignore the high similarity that appears on the cover page and the reference, but not in your report content. You must make sure your submitted report has a similarity lower than 12% from a single source. Otherwise, your report will not be marked.

"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 20 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) contains 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 and include clear sample screenshots with descriptions in your report:

    a. Identify the data types (nominal, ordinal, continuous or discrete) for all attributes. **(10 marks)**
    b. Clean the dataset to remove/patch the missing records/values and record all changes. Explain why you chose such a strategy for the data treatment. Conclude how may tuples are affected and reveal the number of tuples in the dataset before/after the treatment. **(20 marks)**

    [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. Shuffle the dataset to prevent unexpected bias in the partition phase. **(5 marks)**
   b. Partition the dataset into Training and Test sets with a 7:3 ratio. List a sample dataset with headers from both datasets in your report. **(5 marks)**
   c. Pick any five (5) attributes and use visualisation methods to present them with descriptions. **(30 marks, 6 marks for each attribute)**

   [40 marks]

3. Before moving into any prediction model, the company wants to know which store sold the most breads and which store sold the least breads. Use a bar chart to present the revenue from selling breads for the outlets. Based on the char you created, explain what you see and answer the question from the company.
   [10 marks]

4. A senior data scientist raises a concern regarding the correctness of the collected data. Your team is required to verify the data correctness by observing the data with controlled criteria. Report your findings on whether there is any suspicious thing you observed from the data with screenshots as the evidence.
   [15 marks]

5. Present all your answers in the form of a high-quality written report. Create a table showing the distribution of work done by each team member on the assignment.
   [5 marks]

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 item should be prepared in the pdf format, named ***COS10022_("Your Group Number")*** and submitted:

- One (1) piece of a written report of not more than 20-page long with the Assignment Cover Sheet, which should be digitally signed by all group members.

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

Rubric: Subtask 1

Marks	Data Observation and Cleaning (30 marks)
10.0	Correctly identify the data types of all attributes.
20.0	The strategy of data treatment is reasonable, the reasons for choosing such strategies are revealed. The changes to the dataset after the data cleaning process is clearly revealed and compared with the original data.

Rubric: Subtask 2

Marks	Data Preparation (40 marks)
5.0	How the shuffling is processed is clearly revealed and the dataset is properly shuffled.
5.0	The dataset partitioning is executed with no issues.
30.0	The five selected attributes are visualised, and the distributions of the values are described in the report.

Rubric: Subtask 3

Marks	Conventional Data Analysis (10 marks)
5.0	The bar chart is correctly presented with the desired information.
5.0	The question from the company is completely answered.

Rubric: Subtask 4

Marks	Data Verification (15 marks)
15.0	The potential issue is spotted from the verification.

Rubric: Subtask 5

Marks	Quality of the Written Report (5 marks)
5.0	The report is written in a professional way and the table of work distribution to all group members is included in the report.

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