**Assessment Cover Sheet**

This Assessment Cover Sheet is only to be attached to

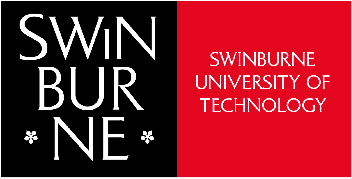
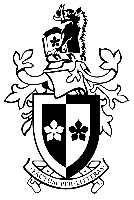
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ASSESSMENT DETAILS** | | | | | | |
| Unit title | | Introduction to Data Science | Tutorial /Lab Group |  | | Office use only |
| Unit code | | COS30045 | Due date | 12 Oct 2020 | |  |
| Name of lecturer/tutor | | Vong Wan Tze | | | |  |
| Assignment title | | Framing a Business Problem into a Data Analytic Problem | | | | Faculty or school date stamp |
| **STUDENT(S) DETAILS** | | | | | | |
| Student Name(s) | | | | | Student ID Number(s) | |
| (1) |  | | | |  | |
| (2) |  | | | |  | |
| (3) |  | | | |  | |
| (4) |  | | | |  | |

hard copy submission of assessments.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **DECLARATION AND STATEMENT OF AUTHORSHIP** | | | | | |
| 1. I/we have not impersonated or allowed myself/ourselves to be impersonated by any person for the purposes of this assessment. 2. This assessment is my/our original work and no part of it has been copied from any other source except where due acknowledgement is made. 3. No part of this assessment has been written for me/us by any other person except where such collaboration has been authorised by the lecturer/tutor concerned. 4. I/we have not previously submitted this work for this or any other course/unit. 5. I/we give permission for my/our assessment response to be reproduced, communicated, compared and archived for plagiarism detection, benchmarking or educational purposes.   I/we understand that:   1. Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to exclusion from the University. Plagiarised material can be drawn from, and presented in, written, graphic and visual form, including electronic data and oral presentations. Plagiarism occurs when the origin of the material used is not appropriately cited. | | | |
| **Student signature/s**  I/we declare that I/we have read and understood the declaration and statement of authorship. | | | |
| (1) |  | (2) |  | |
| (3) |  | (4) |  | |

Further information relating to the penalties for plagiarism, which range from a formal caution to expulsion from the University is contained on the Current Students website at **https://www.swinburne.edu.my/current-students/manage-course/exams-results-assessment**

Copies of this form can be downloaded from the Student Forms web page at **https://www.swinburne.edu.my/current-students/manage-course/exams-results-assessment/how-to-submit-work.php**



**COS10022**

**Introduction to Data Science**

Semester 2, 2020

Assessment Title

**Framing a Business Problem into a Data Analytic Problem**

|  |  |  |
| --- | --- | --- |
| **GROUP NO** |  |  |
|  |  |  |
| **GROUP MEMBERS** |  | (1)  (2)  (3)  (4) |

# Write a short description of the purpose of this section: E.g. In Phase 1 of the Data Analytics Lifecycle, a data science team learns the business domain, assesses the resources available and formulate initial hypotheses (IHs) to test and begin learning the data. The file *BigMart.csv* contains a dataset of 1,559 products across 10 stores in different cities. The following steps were taken to formulate hypotheses that can be tested with the *BigMart* dataset:

# Data Structure

* Identify the dependent variable(s) and the independent variable(s); (Practice 4.1: Titanic Survival Prediction)
* Determine the data types (nominal, ordinal, continuous or discrete) of the dependent and independent variables;

# Data Exploration

* Discuss the relevance of each independent variable for the prediction of the dependent variable; (Do literature search…use articles to support your discussion)
* E.g. Lu Yi et al. (2015) reported that consumers must have active product engagement in a retail store in order to increase the potential of them purchasing a product. Increasing the visual access and visual exposure of a product contributes to higher purchasing opportunities by the consumers.

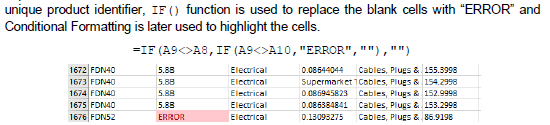
# Hypothesis Generation

* Develop five (5) store level or product level hypotheses based on the discussion in Step 1(b);
* E.G., “Stores located in city areas have higher sales because of higher demand for household products”.
* E.g. “Increase visibility of product …..increase sales..”

# Write a short description of the purpose of this section: E.g. In Phases 2 and 3 of the Data Analytics Lifecycle, the data science team assess the quality of the dataset, learns the relationships between variables and subsequently selects key variables and the most suitable model based on the goal of a project. The following steps were taken to prepare the datasets for model building.

# (a) Data Cleaning

* Make a copy of the sample worksheet in the BigMart.csv file and rename the new worksheet as pre-processing.
* Impute variables with missing or invalid data on the pre-processing worksheet. Explain the imputation strategies and steps involved.
* Explain the Excel formula used for missing data imputation and provide a screen capture of the resulting output. An example is given below:



# (b) Data Transformation

* 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. (Task 4.1: Titanic survival prediction case study)
* Make a copy of the pre-processing worksheet and rename the new worksheet as train.
* Explain the Excel formula used for data transformation and provide a screen capture of the resulting output.

# (c) Train-Test Split

* Produce a test dataset from the training dataset and rename the new worksheet as test. Explain the strategies and steps involved in this step.
* 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’).
* Explain the Excel formula used to randomise the data and split the dataset into training and test set. Provide also screen captures of the resulting outputs.

1. **Model Planning**

* 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.
* Identify an analytical model (such as regression, classification, clustering etc.) suitable for solving the creditability prediction problem. Defend your answer with an explanation.

1. Write a short description of the purpose of this section: E.g. The data science team aims to build a supervised model to predict the sales of a product. To understand the usefulness of the predictive model, a literature search was performed to identify the business values that can be acquired from the ability to automatically predict the expected sale of a product.

**(a) Business Value 1** (e.g. optimising product placement in retail….)

**(b) Business Value 2**

**(c) Business Value 3**

# Distribution of Work

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Student Number** | **% Contribution** |
| Name 1 | Student Name 1 | e.g. 25% |
| Name 2 | Student Name 2 | e.g. 25% |
| Name 3 | Student Name 3 | e.g. 25% |
| Name 4 | Student Name 4 | e.g. 25% |

# References

Ensure that you include in-text citation in the paragraphs and provide the reference list here