Group Assessment: This assessment can be completed individually or in groups of 3 people. **Students are responsible for the formation of groups, or students will be randomly assigned to groups in Week 5.**

Overview

This data analytics assessment requires you to analyse a data set, identify data insights, build and evaluate a number of classification models, evaluate the results, and make recommendations based on your work.

Assessment Groups & Problem Set

Each group will work on a particular problem set. This will consist of high-level details of the problem, meta-data (high-level data descriptions), and a data set. See the sections below for more details of each Problem Set and Dataset to use.

Each group will work on only one of these Problem Sets. This is based on your assigned Team Number (see Sign-In sheet).

Problem Set 1: For teams whose numbers end in 1, 3, 5, 7, 9 (odd numbers)
Problem Set 2: For teams whose numbers end in 2, 4, 6, 8, 0 (even numbers)

You must complete the Problem Set for your Group Number.
If you complete the Wrong Problem Set, this will result in a Zero mark.

Format of Submission

You will submit a Notebook containing all your work, Python code, comments, and MarkDown descriptions of your work, your findings and recommendations.

All work should be in the notebook and completed using Python.
You do not have to include every line of code you write. The notebook should only contain the code relevant to your analysis and what you want to convey about your work and findings i.e. the story you want to give.

A template notebook is available on BrightSpace. You can use this template, which has suggested sections.

Only One student should submit the assessment i.e. only one submission.

Marking Scheme

Marks will be awarded based on the work submitted for each section and based on the **depth of detail**, data insights, explanations, documentation and recommendations given and documented in the notebook.

The marking scheme is divided into the following:

- 5% Problem Definition and setup
- 20% Data Exploration and Findings
- 20% Details of any data preparation, data enrichment, feature engineering, feature reduction, etc
- 20% Details of classification models created, explorations of these, etc
- 20% Details of evaluation and performance measures and recommendations from these
- 15% Discussion of work completed and recommendations

All sections listed above should be documented using MarkDown Cells and code comments. This should document your work, what you did, why you did it, what you learned, how this feeds into and from other parts of your notebook, etc.

For each bullet point above, Marks are allocated, on a sliding scale, based on the depth of detail given, data insights identified, explanations of these and outcomes, comments/documentation of code, etc.

For example, if 20% are available,

	0 marks	1-3 Marks	4-10 marks	11-15 marks	16-20 marks
Marking area	due to lack of code, explanations etc. e.g. just running the Pandas Profiling report, etc	minimal work, explanations/details e.g. just running the Pandas Profiling report with a few general comments. Unnecessary code. Little or no discussion	Some level of analysis with supporting relevant code. No unnecessary code. Good explanations. Well structured, logical flow of code and explanations. Data insights.	Good code, comments, explanations, and insights. No unnecessary code. Clear structure. Some presentation of learning and outcomes	Excellent/exceptional code, comments, explanations, insights, and a clear demonstration of implications. Detailed analysis of outcomes. Connections to previous and next sections

Important: Some additional research might/will be needed on the various Python libraries/packages/functions, how the algorithms work, good programming practice, etc.

The notebook for your assessment must contain the name, student number, and course code (**TU??**) for each student in the group. **Failure to give this information will incur a 10% penalty.**

There will be a 10% penalty deduction will be applied for each day, or part thereof, the assessment is late.
There is no penalty for submitting early.

Plagiarism

Each submission must be original work as plagiarism will result in a **zero** mark (0%). Each student is responsible for ensuring they are complying with the General Assessment Regulations and the TU Dublin Plagiarism Policy

TU Dublin Plagiarism Policy: https://tudublin.libguides.com/c.php?g=674049&p=4794713
https://www.tudublinsu.ie/advice/exams/breachesofregulations/
See the Student Handbook for your course for additional details about plagiarism

Ensure you are compliant with TU Dublin, and the Faculty of Computing, Digital and Data policy on usage of Large Language Models (LLMs), (e.g. ChatGPT, etc) and other similar tools and software. The usage of LLMs is not permitted for this assessment.

Team Work

Good teamwork is vital throughout this assessment. Everyone should contribute equally. Teams should meet regularly online, discuss the assessment, divide up tasks, perform tasks and experimentation, discuss results and outcomes, and prepare the submission of the assessment. Having a difference of opinion is common. Teamwork involves discussion, equal contributions and compromises. The aim is to improve the overall outcome of the work.

Teamwork can be challenging, and it is vital for everyone to play their part in a fair and equal manner.

Assessment Feedback

I will endeavour to mark the assessments and provide feedback via Brightspace VLE, within two to three (working) weeks of the assessment submission date. This will consist of a mark and a short comment on the assessment. This comment will include areas where you did well and areas where you lost marks. This will be provided to the student who submitted the assessment in Brightspace. That student should share the mark and feedback with other members of the group.

Problem Set 1 – Portuguese Banking Marking Campaign

The data set and problem are from a well-known (and discussed) use-case and the data set is available for download from the UCI ML Repository. This website contains additional information about the data set.

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

The data set is related to a direct marketing campaign for a Portuguese banking institution. The bank conducts marketing campaigns and uses their call center to contact their customers using phone calls. The purpose of this project is to identify customers who are most likely to subscribe to a term deposit account based on previous marketing campaigns.

Data Set

The data sets provided contain the results of the previous marketing campaigns. The website (given above) contains the data files and attribute descriptions of the data set.

NOTE : There are two versions of the data set given/ One of the data sets contains 17 input attributes. The second data set contains 20 input attributes. It is this data set (20 attributes) that you should use for the assessment. The files for this data set are called 'bank-additional' (see bank-additional-full.zip).

The data sets contain two files. the 'bank-additional.csv' contains a 10% sample (4,119 records) of the entire data set. Do not use this file. Instead, use the entire data set included in the file called 'bank-additional-full.csv' which contains **41,188 records**. Use this data set.

Published Paper

A published paper is available that describes the project, the data set and what data mining methods were applied. Here is the full citation for the paper and a link to where you can find it. You should study and read this paper carefully.

S. Moro, P. Cortez and P. Rita. *A Data-Driven Approach to Predict the Success of Bank Telemarketing*, Decision Support Systems, Elsevier, 62:22-31, June 2014

You can download the paper using this link https://www.dropbox.com/s/c4pimj8aj2f0pkn/Bank-Research-Paper.pdf?dl=1

You are required to build several classification models using Python, evaluate the results and make recommendations.

Problem Set 2 – Supermarket Marking Campaign

A supermarket is beginning to offer a new line of organic products. The supermarket's management would like to determine which customers are likely to purchase these products.

The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of their loyalty program participants and has now collected data that includes whether or not these customers have purchased any of the organic products.

The **ORGANICS** data set contains approx. 20,000 observations and 18 variables. The variables in the data set are shown below with the appropriate roles and levels. For this assessment, the target variable is the binary variable **ORGYN**. The data set will be available from within your SAS account.

Data Set

Name	Model Role	Measurement Level	Description
CUSTID	ID	Nominal	Customer loyalty identification number
GENDER	Input	Nominal	M = male, F = female, U = unknown
DOB	Input	Interval	Date of birth
EDATE	Input	Unary	Date extracted from the daily sales data base
AGE	Input	Interval	Age, in years
AGEGRP1	Input	Nominal	Age group 1
AGEGRP2	Input	Nominal	Age group 2
TV_REG	Input	Nominal	Television region
NGROUP	Input	Nominal	Neighborhood group
NEIGHBORHOOD	Input	Nominal	Type of residential neighborhood
LCDATE	Input	Interval	Loyalty card application date
LTIME	Input	Interval	Time as loyalty card member
ORGANICS	Input	Interval	Number of organic products purchased
BILL	Input	Interval	Total amount spent
REGION	Input	Nominal	Geographic region
CLASS	Input	Nominal	Customer loyalty status: tin, silver, gold, or platinum
ORGYN	Input	Binary	Organics purchased? 1 = Yes, 0 = No
AFFL	Input	Interval	Affluence grade on a scale from 1 to 30

The data set, in CSV format, can be downloaded using this link
https://drive.google.com/file/d/1jF8X9oj5ZGpJoP0uDTRh8zz2ldRcexHM/view?usp=drive_link

You are required to build several classification models using Python, evaluate the results and make recommendations.