

Course: MSc Data Analytics

Module Code: CO4762

Module Title: Knowledge Discovery

Title of the Brief:

Type of assessment: [report, portfolio]

This assessment is worth 50% of the overall module mark.

This Assessment Pack consists of a detailed assignment brief, guidance on what you need to prepare, and information on how class sessions support your ability to complete successfully. You'll also find information on this page to guide you on how, where, and when to submit. If you need additional support, please make a note of the services detailed in this document.

How, when, and where to submit:

The assessment release date is: 23/01/2025.

The assessment deadline date and time is:

Submission deadline:

28/03/2025 @23:59

Presentation date:

TBD

Feedback will be provided by: 19/04/2024.

You should aim to submit your assessment in advance of the deadline.

Note: If you have any valid mitigating circumstances that mean you cannot meet an assessment submission deadline and you wish to request an extension, you will need to apply online, via [MyUCLan](#) with your evidence **prior to the deadline**. Further information on Mitigating Circumstances via [this link](#).

SUBMISSION DETAILS

You are required to submit the following:

1. Add the completed assignment coversheet (see page 4)
2. All deliverables mentioned in section [Deliverables] of the **Detailed Assignment Brief**

We wish you all success in completing your assessment. Read this guidance carefully, and if any questions, please discuss with your Module Leader.

Additional Support available:

All links are available through the online [Student Hub](#)

1. Academic support for this assessment will be provided by contacting **Panayiotis Andreou** pgandreou@uclan.ac.uk.
2. Our **Library resources** link can be found in the [library area](#) of the Student Hub or via your subject librarian at CyprusLibrary@uclan.ac.uk
3. For help with Turnitin, see [Blackboard and Turnitin Support](#) on the Student Hub
4. If you have a disability, specific learning difficulty, long-term health or mental health condition, and not yet advised us, or would like to review your support, **Student Support Officers (cyprusstudentsupport@uclancypus.ac.cy)** can assist with reasonable adjustments and support. To find out more, you can visit the **Student Support Service** page of the [Student Hub](#). You can also call +357 24694026 or +357 24694108 or +357 24694073.
5. For mental health and wellbeing support, please complete our [online referral form](#) or email the Psychological Wellbeing and Counselling Centre at UCLan Cyprus at wellbeing@uclancypus.ac.cy.
6. For any other support queries, please contact **Student Support** via CyprusStudentSupport@uclan.ac.uk.
7. For consideration of Academic Integrity, please refer to detailed guidelines in our [policy document](#). All assessed work should be genuinely your own work, and all resources fully cited.
8. For advice on the use of Artificial Intelligence, please refer to [Categories of AI tools](#) guidance.

For this assignment, you are not permitted to use any category of AI tools.

Preparing for your assignment.

Refer to the Module Information Pack to understand the Learning Outcomes and Marking Criteria.

Feedback Guidance:

Reflecting on Feedback: how to improve.

From the feedback you receive, you should understand:

- The grade you achieved
- The best features of your work
- Areas you may not have fully understood
- Areas you are doing well but could develop your understanding.
- What you can do to improve in the future - feedforward

Next Steps:

- List the steps have you taken to respond to previous feedback.
- Summarise your achievements
- Evaluate where you need to improve here (keep handy for future work):

Coursework Cover Sheet

Students should complete the input fields contained in this form and attach it in front of your formal assessment submission. All fields within this form are required. Please ensure that check boxes and radio buttons are appropriately selected. The last three questions are just for you to personally consider.

Department and assessment information:

School Name: School of Sciences

Assessment title: Knowledge Discovery assignment

Course Title: MSc Data Analytics

Module Title: Knowledge Discovery

Module Code: CO4762

Year of Study: 2023-2024

Academic Misconduct / Plagiarism Declaration

By attaching this front cover sheet to my assessment I confirm and declare that **I am the sole author of this work**, except where otherwise acknowledged by appropriate referencing and citation, and that I have taken all reasonable skill and care to ensure that no other person has been able, or allowed, to copy this work in either paper or electronic form, and that prior to submission I have read, understood and followed the University regulations as outlined in the [Academic Integrity Policy and Procedure for Academic Misconduct](#)

Have you checked the following? This will help your assessment achievement.

I have applied the learning outcomes for this module ☐

I have checked for Academic Integrity via Turn-it-in ☐

I have followed the guidance in the Assessment Brief and have not used AI to boost my grade unfairly. ☐

I have used references in accordance with instructions in the Assessment Brief ☐

I have proofread my work for spelling, grammar and punctuation. ☐

I have checked that the word count/size of this submissions accords with the guidance provided in the Assessment Brief. ☐

Well-being

We wish to support any student who is experiencing mitigating circumstances which prevents students from performing to the best of their ability when completing or submitting

assignments. If you are experiencing such circumstances, then you may apply for Mitigating Circumstances. Wherever possible this must be done prior to handing in your assignment.

Do you need to apply for mitigating circumstances for this assignment Please select Yes / No

Please refer to the [Mitigating Circumstances Policy](#)

Questions you may wish to consider:

1. Have I allowed sufficient time to prepare this assessment?
2. Have I reflected on previous feedback and made improvements in accordance with advice?
3. What grade am I expecting?

CO4762: Knowledge Discovery

Assignment

Date Issued: 23/01/2025

Submission deadline: 28/03/2025

Presentation date: TBD

IMPORTANT

- **Read the marking scheme carefully.**
- **This is an individual project** and no group work is permitted.

I. Purpose

The purpose of this assignment is to allow students to get familiarized with all the phases of predictive modelling. You have been hired by SuperApp, a fictional supermarket company, as a Data Analyst, to assist in setting up their marketing strategy for a new line of products. Your purpose is to analyse existing customer data and discover which customers are likely to purchase these products.

In particular, in this assignment, you will:

- Prepare a dataset for analysis purposes;
- Explore the data and understand the dataset and its main dimensions by highlighting key findings;
- Analyse the data using a range of predictive analytical techniques to reveal important insights and perhaps hidden patterns;
- Create a comprehensive business report that encompasses the key findings of all aforementioned parts.

II. Requirements

SuperApp is a supermarket that is offering a new line of products. The supermarket's management wants to determine which customers are likely to purchase these products. As an initial buyer incentive plan, the supermarket has provided coupons for the new line of products to all of the loyalty program participants and has collected data about whether these customers have purchased any related products recently.

In particular, the management of the supermarket has created a dataset that includes variables about demographics and loyalty status purchase information about products. The variables in the data set are shown below with the default roles and levels.

Name	Description
CustomerID	Customer loyalty identification number
ProsperityClass	Prosperity class on a scale from 1 to 100, 100=highest prosperity class
Age	Customer Age, in years
ResType	Type of residential neighbourhood
Gender	Customer Gender
District	District
TVReg	Television region
CardClass	Loyalty status: Bronze, Silver, Gold, or Platinum
AmountSpent	Total amount spent
CustomerRetention	Total months as a customer
CountProducts	Number of products purchased
Target	Purchased new line of products recently: 1 = Yes, 0 = No

The above dataset, contains more than 22K observations. The data granularity level is customer aggregated, and records are depicted in the form of a consolidated view of all attributes/dimensions from a star schema. These attributes describe customer demographics, customer loyalty and purchases.

You are required to prepare, explore and analyse the data using multiple predictive modelling techniques, and create a business report that will summarize your findings. SAS Enterprise Miner will be used for all aspects of data preparation, exploration and analysis. Microsoft Word will be used for the compilation of the comprehensive report. In particular, you are required to complete the following parts:

Part	Description	Details	Requirements
A	Business Report	A formal business report containing a cover, an executive summary summarizing the results of analysis (i.e., results of parts B,C,D).	

B	Data Preparation and Exploration	A document detailing the steps of the data preparation phase.	Tasks DPE1-DPE7
C	Predictive Modelling	A document describing the analysis of the data using a range of descriptive, predictive and prescriptive analytical techniques, which reveals important insights and perhaps hidden patterns.	Tasks PM1-PM5
D	Model Evaluation	A report that assesses the structure, performance, and resilience of the predictive models used in part C.	Tasks ME1-ME2
E	Presentation	A critical discussion comparing the predictive modelling techniques used in C, including their advantages and disadvantages.	

Tasks for Part B: Data Preparation and Exploration

- DPE1. Create the data source and place it into a new diagram.
- DPE2. Adjust the role and level of each variable. Justify your decisions for each variable.
- DPE3. There are two target variables. Discuss how these can be used for predictive modelling. Discuss if AmountSpent should be used as an input for a model for predicting Target.
- DPE4. Discuss the distribution of the Target variable. Provide insight on the correlations of target with other attributes.
- DPE5. Attach the StatExplore tool to the data source. Discuss the results with regards to Missing Values and Imputation.
- DPE6. Partition the data source for Training 50% and Validation 50%.

Tasks for Part C: Predictive Modelling

During this task you are requested to create a number of predictive models for predicting the Target attribute and assess their prediction accuracy.

For each predictive model, you will need to discuss the following elements:

- i. Special data preparation requirements of the model
- ii. Prediction accuracy of the model
- iii. Interpretation of results of the model

Your analysis should include at least 3 models of each of the families listed below:

- PM1. **Decision Trees**
- PM2. **Regressions**
- PM3. **Clustering**
- PM4. **Neural Networks**
- PM5. **Support Vector Machines**

Tasks for Part D: Model Evaluation and Scoring

- ME1. Using **Model Comparison**, evaluate the predictive models with regards to Misclassification Rate. Use the ROC curve to demonstrate which predictive model is the best.
- ME2. Use the model that was selected in the previous step to **Score** a fresh copy of the data source. Confirm the accuracy of the prediction.

Task for Part E: Presentation

You should prepare a presentation that critically presents the results. Your presentation **must** (a) include a description of the historical data; (b) describe and interpret the most accurate decision tree; (c) describe and interpret another modelling technique; (d) explain the cut - off point business wise; (e) describe the gain of using predictive modelling (cumulative lift) of the selected model.

Additionally, you will need to prepare one/two slides of conclusions/recommendations, focusing on which customers will buy the new line of products to present to the management team.

Total presentation time: 10 minutes + 5 minutes of questions

III. Deliverables

You are required to produce one deliverable as described below.

Part	Description	Marking Range	Deliverable
A	Business Report	0-15	<ol style="list-style-type: none"> 1. A word document (.docx) report that includes the output of parts A-D 2. An XML file with the complete SAS Enterprise Miner project 3. A PowerPoint presentation (.pptx) for part E
B	Data Preparation and Exploration	0-10	
C	Predictive Modelling	0-50	
D	Model Evaluation and Scoring	0-10	
E	Presentation	0-15	

IMPORTANT: The requirements provided in the previous section may not be sufficiently defined. During the specifications, you will need to record your assumptions and how these have influenced your report design.

IV. Grading Criteria

Marks will be awarded based on the following criteria. In assessing the work within a section, factors such as simplicity, quality and appropriateness of comments, and quality and completeness of the design will be considered.

Part	Description	Criteria
A	Business Report	<p><u>Cover</u></p> <ul style="list-style-type: none"> • 0 – No attempt • 1 – Poor cover • 2 – Professional cover <p><u>Structure</u></p> <ul style="list-style-type: none"> • 0 – No attempt • 1 – Uses formatted headings • 1 – Provides TOC automatically generated from headings • 1 – Develops Header and Footer • 2 – Provides an introduction to each section <p><u>Executive Summary</u></p> <ul style="list-style-type: none"> • 0 – No attempt • 3 – Clarity: Use definite, specific, concrete language to discuss your findings • 2 – Conciseness: Summarize the key findings omitting needless analysis • 2 – Coherency: Information elements should hold together so that progress from one point to the other seems inevitable
B	Data Preparation and Exploration	<p><u>DPE1</u></p> <p>0 – No attempt or task performed incorrectly</p> <p>1 – Task performed correctly with no deficiencies</p> <p><u>DPE2</u></p> <p>0 – No attempt</p> <p>1 – Justifies decisions for each variable with minor errors</p>

		<p>2 – Justifies decisions for each variable correctly</p> <p><u>DPE3</u></p> <p>0 – No attempt</p> <p>1 – Justifies how Target can be used only</p> <p>2 – Justifies how both Target and AmountSpent can be used</p> <p><u>DPE4</u></p> <p>0 – No attempt or task performed incorrectly</p> <p>1 – Valid observations about distribution of Target</p> <p>2 – Valid observations about distribution of Target and correlations with other attributes</p> <p><u>DPE5</u></p> <p>0 – No attempt or task performed incorrectly</p> <p>1 – Trivial conclusions about exploration</p> <p>2 – Critical discussion about exploration demonstrating knowledge about when Imputation must be used</p> <p><u>DPE6</u></p> <p>0 – No attempt or task performed incorrectly</p> <p>1 – Dataset prepared correctly for training and validation</p>
C	Predictive Modelling	<p>0-10 for each Predictive Model</p> <p><u>Special data preparation requirements</u></p> <p>0 – No attempt or incorrect</p> <p>1 – Lists special data preparation requirements</p> <p>2 – Discusses special data preparation requirements or justifies why they are not required</p>

		<p><u>Development of Predictive Model</u></p> <p>0 – No attempt</p> <p>1 – Develops a correct model with default settings</p> <p>2 – Develops an optimized model but does not justify the tuning of parameters</p> <p>3 – Develops an optimized model, fully justifying the tuning of parameters</p> <p><u>Prediction Accuracy</u></p> <p>0 – No attempt</p> <p>1 – Lists the prediction accuracy of the model</p> <p>2 – Discusses the prediction accuracy of the model using appropriate metrics</p> <p><u>Interpretation</u></p> <p>0 – No attempt or incorrect interpretation</p> <p>1 – Correct interpretation of the model with minor errors</p> <p>2 – Correct interpretation of the model</p> <p>3 – Correct interpretation of the model providing insight using appropriate diagrams</p>
D	Model Evaluation and Scoring	<p><u>Development of Model Comparison</u></p> <p>0 – No attempt or incorrect</p> <p>1 – Developed correctly</p> <p><u>Development of Scoring</u></p> <p>0 – No attempt or incorrect</p> <p>1 – Developed correctly</p> <p><u>Findings/Conclusions 0-8</u></p> <p>0 – No attempt</p>

		<p>2 – Trivial or obvious conclusions</p> <p>4 – Reasonable conclusions that are supported by one evaluation metric</p> <p>6 – Provides evidence of investigative skills for the evaluation of the predictive models using at least two evaluation metrics</p> <p>8 – Effective and concise story-telling, backed up by appropriate evidence and data visualizations</p>
E	Presentation	<p>Presentation will be evaluated on several criteria such as style, structure and flow, engagement, and based on the responses of the interview questions</p> <p>0-15</p>

Submission of assignment work

- Anonymous marking is being used. You may include your University ID number (“G2...”) on the work. Apart from this, avoid doing anything that would allow you to be identified from your work.
- *Keep a complete copy of the work you hand in.*
- Avoid submitting work at the last minute, but if there is a technical problem uploading to Blackboard, email the zip file to me before the deadline and upload the work when Blackboard is available.
- Include the coursework cover sheet (attached at the beginning of this assignment brief) at the beginning of your submission report, filled in accordingly.

Extenuating circumstances, extensions and late work

Except where an extension of the hand-in deadline date has been approved (see https://www.uclan.ac.uk/students/study/examinations_and_awards/extenuating_circumstances.php), work that is handed in up to 5 days late will be capped to 50%. After this, it will receive a mark of 0%:

Cheating

The consequences of cheating in assessments are serious. Cheating is using or attempting to use unfair means to enhance performance. This includes plagiarism (presenting someone else's work as if it was your own), collusion (working with others on an individual assignment), taking prohibited material into examinations and allowing other students to access your work. Make sure that you do not give someone the opportunity to steal your work (e.g. *by asking them to print it out for you*). We tell students about cheating both during induction and in your student handbook, but if you have any doubt about what cheating is or how to reference material properly, please ask a tutor. We recommend that you use the Harvard system for referencing.

The University operates an electronic plagiarism detection service where your work may be uploaded, stored and cross-referenced against other material. The software searches the World Wide Web and extensive databases of reference material to identify duplication.

For more information about plagiarism, please see the University Academic Regulations and the Assessment Handbook (http://www.uclan.ac.uk/aqasu/academic_regulations.php). See the Student Union website: <http://www.uclansu.co.uk/academicmatters/unfairmeans>

Reassessment and Revision

Reassessment in written examinations and coursework is at the discretion of the Course Assessment Board and is dealt with in accordance with University policy and procedures.