



## **CO4762: Knowledge Discovery**

### **Assignment**

**Date Issued:** 04/02/2021

**Hand in Date:** 29/04/2021

### **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"><li>1. A word document (.docx) report that includes the output of parts A-D</li><li>2. An XML file with the complete SAS Enterprise Miner project</li><li>3. A PowerPoint presentation (.pptx) for part E</li></ol>
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	<u>Cover</u> <ul style="list-style-type: none"><li>• 0 – No attempt</li><li>• 1 – Poor cover</li><li>• 2 – Professional cover</li></ul> <u>Structure</u> <ul style="list-style-type: none"><li>• 0 – No attempt</li><li>• 1 – Uses formatted headings</li><li>• 1 – Provides TOC automatically generated from headings</li><li>• 1 – Develops Header and Footer</li><li>• 2 – Provides an introduction to each section</li></ul> <u>Executive Summary</u> <ul style="list-style-type: none"><li>• 0 – No attempt</li><li>• 3 – Clarity: Use definite, specific, concrete language to discuss your findings</li><li>• 2 – Conciseness: Summarize the key findings omitting needless analysis</li><li>• 2 – Coherency: Information elements should hold together so that progress from one point to the other seems inevitable</li></ul>

B	Data Preparation and Exploration	<p><u>DPE1</u> 0 – No attempt or task performed incorrectly 1 – Task performed correctly with no deficiencies</p> <p><u>DPE2</u> 0 – No attempt 1 – Justifies decisions for each variable with minor errors 2 – Justifies decisions for each variable correctly</p> <p><u>DPE3</u> 0 – No attempt 1 – Justifies how Target can be used only 2 – Justifies how both Target and AmountSpent can be used</p> <p><u>DPE4</u> 0 – No attempt or task performed incorrectly 1 – Valid observations about distribution of Target 1 – Valid observations about distribution of Target and correlations with other attributes</p> <p><u>DPE5</u> 0 – No attempt or task performed incorrectly 1 – Trivial conclusions about exploration 2 – Critical discussion about exploration demonstrating knowledge about when Imputation must be used</p> <p><u>DPE6</u> 0 – No attempt or task performed incorrectly 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> 0 – No attempt or incorrect 1 – Lists special data preparation requirements 2 – Discusses special data preparation requirements or justifies why they are not required</p> <p><u>Development of Predictive Model</u> 0 – No attempt 1 – Develops a correct model with default settings 2 – Develops an optimized model but does not justify the tuning of parameters 3 – Develops an optimized model, fully justifying the tuning of parameters</p> <p><u>Prediction Accuracy</u> 0 – No attempt 1 – Lists the prediction accuracy of the model 2 – Discusses the prediction accuracy of the model using appropriate metrics</p>

		<u>Interpretation</u> 0 – No attempt or incorrect interpretation 1 – Correct interpretation of the model with minor errors 2 – Correct interpretation of the model 3 – Correct interpretation of the model providing insight using appropriate diagrams
D	Model Evaluation and Scoring	<u>Development of Model Comparison</u> 0 – No attempt or incorrect 1 – Developed correctly  <u>Development of Scoring</u> 0 – No attempt or incorrect 1 – Developed correctly  <u>Findings/Conclusions 0-8</u> 0 – No attempt 2 – Trivial or obvious conclusions 4 – Reasonable conclusions that are supported by one evaluation metric 6 – Provides evidence of investigative skills for the evaluation of the predictive models using at least two evaluation metrics 8 – Effective and concise story-telling, backed up by appropriate evidence and data visualizations
E	Presentation	Presentation will be evaluated n several criteria such as style, structure and flow, engagement, and based on the responses of the interview questions  0-15

### 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.

### 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](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](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.