ROBERT GORDON UNIVERSITY ABERDEEN

COURSEWORK OVERVIEW

Module Number & Title	**CMM535** **Data Science Development**
Coursework Part 2	This is Part 2 of a two-part coursework for the module and is detailed in this brief. This Part will be released on Friday 18 March This part is an **INDIVIDUAL EXERCISE**
Submission Method	Coursework must be submitted electronically via the designated (CMM535 Coursework Part 2) Assessment Dropbox on CampusMoodle.
Deadline	PART 1: Friday 11 March 2022 @ 16:00 PART 2: Friday 29 April 2022 @ 16:00 **The submission deadline is 4pm** – whilst a 30-minute grace period has been added for technical issues, you should aim to submit by 4pm.
Module Co-ordinator	Module Coordinator details – Ines Arana/David Lonie Email: d.p.lonie@rgu.ac.uk For Staff Office Hours, visit: http://campusmoodle.rgu.ac.uk/course/view.php?id=96625
LOs Assessed	This part of the coursework assesses the following learning outcomes of this module: **LO 1** Discuss the main concepts and tools for a data science project. **LO 2** Load, explore, model and visualise data using off-the-shelf tools and packages. **LO 3** Report data science results to a wider audience by tailoring them at different levels of detail. **LO 4** Design, implement and evaluate a data science product that addresses a given data problem taking in account relevant professional, legal, ethical, security and social issues.

AUXILIARY INFORMATION

Please refer to the coursework brief below for detailed information on your submission, including any software packages/versions that should be used, word counts and limits.

Coursework received after the submission deadline indicated above will be regarded as a non-submission (NS) and one of your assessment opportunities will be lost. Coursework extended due to extenuating circumstances shall be assessed in the normal way.

If the **word count** of an assessment is considered critical, then this will be reflected in the assessment criteria for that assessment together with any consequent penalties.

In line with the RGU Assessment Policy, **this coursework has been moderated** by the School of Computing Moderation Panel, which comprised a detailed technical review and panel overview to schedule staggered submissions.

ACADEMIC INTEGRITY

For this assessment, all work is expected to be completed by yourself. You should be able to complete this assessment using the module materials, and your own understanding of the material. If you do consult or re-use any external resources then before submitting this assignment, you should check your submission to ensure that it complies with Academic Regulation A3-2 Student Conduct Procedure and Academic Integrity, including but not limited to:

- *all material identified as originally from a previously published source has been properly attributed by the inclusion of an appropriate citation at the point of use in the text;*
- *direct quotations are marked as such (using "quotation marks" at the beginning and end of the selected text), and*
- *full details of the reference citations have been included in the list of references.*

EXTENSIONS AND DEFERRALS

The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so. For information about extensions and deferrals, please familiarise yourself with the extenuating circumstances defined therein.

INSTRUCTIONS:

This part of the CMM535 coursework requires you to perform a set of Supervised Learning Experiments on a dataset, and also to create a basic deployment of one of the resulting models.
The dataset is available on Moodle and is named **escapesClean.csv**
The experiments are described below, each with a target variable. You will need to select and specify your performance metric(s) appropriate to the nature of the modelling task.

Experiment 1 is a classification task, described below as Model 1 and Model 2. The aim is to compare the performance of two classification models, logistic regression and one other of your choice from the models we have covered in this module, and to compare their performance.

Experiment 2 is a regression task, described below as Model 3 and Model 4. The aim is to compare the performance of two regression models, linear regression and one other of your choice from the models we have covered in this module, and to compare their performance.

Wherever appropriate, experiments should involve cross-validation or bootstrapping to estimate performance metrics. As well as comparing different algorithms, you should try to achieve optimum performance for the models. For example: through feature selection, or pre-processing, prior to running the algorithm; and through tuning the parameters of the algorithm.

You are also required to create a basic deployment of one model using an **R Shiny** file. We shall cover this in Lab 10 of the module.

Once completed you are required to submit **two things** to CampusMoodle**:**
The first file is a **zipped folder containing four files**:

- a **Rmd** markdown file or **R** script containing your R code and relevant text commentary on model fitting and the results,
- a copy of the data file **escapesClean.csv**
- an **rds** file named **model.rds** as specified in the instructions that follow
- a shiny app file **app.R** that deploys the prediction model as described in instructions below

The second file you must submit is an **HTML** format file created by **knitting** your **Rmd** or **R file.** This must not be zipped, as it will be processed through Turnitin on Moodle.

Professional presentation is part of what is being assessed, so submissions missing any of the required files will receive a limited grade (as defined in the marking criteria).

Text comments in the markdown file should be concise and relevant and included as markdown text or code comments. Irrelevant or excessive comments may be penalised.

THE DATASET:

escapesClean.csv has the same context as that from Part 1 of the coursework. However **escapesClean.csv** is a heavily cleaned version of the files **escape.csv** that you used in Part 1 of the coursework. So the context of the underlying data remains the same as in Part 1 in that it relates to fish farming data. As well as having been cleaned, **escapesClean.csv** also has several new variables compared to escapes.csv. You should use only **escapesClean.csv** for this Part of the coursework.
You should download the completed version of **escapesClean.csv** files from the CMM535 Moodle page, save it in a convenient folder, and load the data into RStudio using the read.csv() function.
You should work with and submit a **single Rmd** file for all 4 models specified below.

It is highly recommended to save all relevant files (your R code, the data file, etc) in a single folder, as you will need to upload that folder as part of your submission.
The variables in the dataset are described in an appendix at the end of this document.

MODEL 1: LOGISTIC REGRESSION

Design and implement a Logistic Regression model to predict **Cause**

MODEL 2: CLASSIFICATION MODEL OF YOUR CHOICE

Design and implement another classification model that has been covered in the CMM535 module to predict **Cause**.
Critically compare and contrast the effectiveness of model 1 and model 2 [**Word limit of 150 words**].

MODEL 3: LINEAR REGRESSION

Select data features that will be suitable and relevant for predicting **Number**
Design and implement a Linear Regression model to predict **Number**

MODEL 4: REGRESSION MODEL OF YOUR CHOICE

Design and implement a second Regression model, using the same set of data features as for Model 3, using techniques that have been covered in CMM535 module, to predict **Number**.

Critically compare and contrast the effectiveness of model 3 and model 4 [**Word limit of 150 words**].

MODEL DEPLOYMENT: A BASIC SHINY APP

For whichever model you consider to be best between Model 1 or Model 2, and refitting your chosen the model if necessary to use at most **6** of the most effective predictors for **Cause**, save the model as an **rds** file called **model.rds** using the **saveRDS** function in R.

Create a basic Shiny app with inputs for the values of the predictors in your chosen model, a button labelled *predict* and an output field for the target variable prediction. Create the appropriate code such that when the user enters values for all predictors and click the button, the model is used to display the prediction in the output field.

This may sound daunting, but the examples studied in **Lab 10** can be used as templates that should make this straightforward if you have done that lab exercise.

The app interface should also have a text field summarising the legal basis upon which the original data is being used. You may refer to the original source of the data http://aquaculture.scotland.gov.uk/data/data.aspx for relevant information in this regard. You should also address any additional ethical or social issues that are appropriate for the use of the predictions being generated. [**Word limit of 150 words**]

COURSEWORK PART 2 MARKING

Part 2 of the CMM535 Coursework (as specified in this document) will be assigned a grade **A-F** based on the following criteria:

GRADE	A	B	C	D	E	F
DEFINITION	EXCELLENT Outstanding Performance	COMMENDABLE Meritorious Performance	GOOD Highly Competent Performance	SATISFACTORY Competent Performance	BORDERLINE FAIL Open To Compensation	FAIL Non-Submission or Unsatisfactory
MODEL SELECTION AND IMPLEMENTATION	All the models applied are appropriate. A professional overview of the required steps is presented (including data processing, data modelling and model tuning).	Most of the models applied are appropriate. A very good overview of the required steps is presented (including data processing, data modelling and model tuning).	Most of the models applied are appropriate. A good overview of the required steps is presented (including most of the expected data processing, data modelling and model tuning).	Some of the models applied are appropriate. An adequate overview of the required steps is presented (including some of the expected data processing, data modelling and model tuning).	The model selection or implementation has significant issues in terms of accuracy or completeness.	Little or no modelling implemented, or is flawed to an extent that makes it irrelevant.
EVALUATION	The evaluation of all models is complete and insightful.	The evaluation of most models is complete and insightful. Some evaluation of all models is presented.	The evaluation of most models shows some insight. Some evaluation of most models is presented.	The evaluation of models shows some insight.	The model evaluation has significant issues in terms of accuracy or completeness.	Little or no evaluation, or is flawed to an extent that makes it irrelevant.
PRESENTATION	Professionally presented, insightful narrative demonstrating excellent understanding of assessed concepts.	Very well presented, narrative demonstrating very good understanding of assessed concepts.	Well presented, narrative demonstrating good understanding of assessed concepts.	Adequately presented, narrative demonstrating some understanding of assessed concepts.	Inadequate standard of narrative, or demonstrating very little understanding of assessed concepts.	Very poor standard of narrative, and demonstrating very little or no understanding of assessed concepts.
MODEL DEPLOYMENT	Fully functional Shiny app generating predictions based on Model 4 and a professional LESP summary.	Functional Shiny app generating predictions based on Model 4 and a comprehensive LESP summary.	Functional Shiny app generating predictions based on Model 4 and a good LESP summary.	Shiny app generating predictions based on Model 4 and a LESP summary.	Shiny app partially complete or generating flawed prediction, poor or missing LESP summary.	Shiny app missing or non-functional, missing LESP summary.

OVERALL CM535 MODULE GRADE

The grades from Part 1 and this Part of the CMM535 assessment are each awarded a grade A -F. These are combined as follows to give the overall CMM535 module grade. The **minimum requirements** for each overall grade are as follows:

- Grade A: 2 As.
- Grade B: 1B + 1C.
- Grade C: 1C + 1D.
- Grade D: 1D + 1E.
- Grade E: 1E + 1F
- Grade F: where none of the above requirements are met.

APPENDIX - DATASET DESCRIPTION

The dataset **escapesCleaned.csv** consists of **221** complete rows of data for the following variables:

Season	a category with 4 levels, {Spring, Summer, Autumn, Winter}
Species	a category with 4 levels, describing a type of fish
Age	the average age of fish that escaped
Average.Weight	the average weight in g of escaped fish
Number	the estimated number of escaped fish
Cause	a category with 2 levels, describing the cause of the escape, whether **Human** error or else a **Natural** event
Producing	a category with 2 levels, describing whether the site was producing within last 3 years
SLR	a measure of sea-lice residue recorded at the site
Cu	a measure of level of copper compounds detected in water at the site
Zn	a measure of level of zinc compounds detected in water at the site
N	a measure of level of nitrogen compounds detected in water at the site
P	a measure of level of phosphorus compounds detected in water at the site
Org	a measure of level of organic compounds detected in water at the site