DEPARTMENT OF ACTUARIAL MATHEMATICS AND STATISTICS

F71RA Machine Learning for Risk and Insurance 1
2024-2025 Semester 1 Project

Question	Marks
Part 1	15
Part 2	25
Part 3	30
Part 4	30
Total Marks	100

Attempt **ALL questions**.

You must **work on this project individually**.

You should create **one report (pdf file) with R script in your submission**.

Release Date: Tuesday 8th October from 11am
Submission Date: Friday 22nd November by 4pm
Note: standard penalties apply thereafter for unjustified late submission
Submission process: Canvas Turnitin
Submission Requirements: Report, R code and analysis.
Report Length: maximum of 25 pages (not including appendix of code)
Feedback on this assessmentwill be provided within 2 weeks of submission.

GUIDANCE FOR SUBMISSION OF REPORT:

- Submission will be made through Heriot Watt Canvas Online Submission.
- Submit all files in a .zip file labeled by student numbers e.g. each student "H11111.zip"
- The .zip file should contain your report as a pdf, your R code and data.

1 Instructions

1.1 Marking.

This assignment will be marked out of 100 and will carry a 50% weight in the final mark for the course.

1.2 Submission Deadline.

The deadline for submission of the assignment is 4pm on the Friday 22nd November, 2024 by 4pm. Please make regular back-ups of your work and do not leave final report creation to the last minute. Take careful notice of the expected format, file naming convention required and the file types requested.

The mark for coursework submitted late, but within 5 working days of the coursework deadline, will be reduced by 30%. Coursework submitted more than 5 working days after the deadline will not be marked. In a case where a student submits coursework up to five working days late, and the student has valid mitigating circumstances, the mitigating circumstances policy will apply. Students should be advised in such cases to submit a Mitigating Circumstances form for consideration by the Mitigating Circumstances Committee.

1.3 Form of solutions.

Your solution to the assignment should take the form of a report. It should be prepared using LaTex, Word or other word-processing software. Relevant computer code must be included and should be placed in an Appendix. You may also wish to put some of the graphs that illustrate your solution to Section B in an Appendix.

1.4 Length.

Reports should not exceed 25 pages in length. The page limit includes tables and graphs. Computer code does not count towards your page limit.

1.5 How to submit.

Assignments must be handed in using the coursework turnitin link where two files are expected. You should submit your reports in the following format

1. report in pdf format
2. zip file of the code, data, report. Both file names should be labeled as follows: ReportSTUDENTIDHERE.pdf, FilesSTUDENTIDHERE.zip

1.6 Information to be provided.

Front cover of your assignment (not included in page count) should include the following information (see final page of handout for example):

- the text "'Department of Actuarial Mathematics and Statistics"';
- your full name;
- your matriculation number;
- the degree for which you are you registered;
- the text 'Machine Learning for Risk and Insurance: Project';
- the names of other students that you discussed the assignment with;
- a signed plagiarism declaration following the wording given at the end of this document.

Assignments which omit a signed plagiarism declaration will automatically be awarded zero marks.

1.7 Figures and tables.

These are usually necessary to present the results of empirical work clearly. If you include figures and tables in your report, the discussion of their contents should be contained in the main text and not in captions. Please ensure that figures are easy to interpret and use clear labelling and legends. If you use colour figures, then please print the report in colour.

1.8 Group discussions.

You may discuss this project with your classmates. However, you must conduct your own independent analyses and write your report independently of other students in your class.

1.9 Plagiarism and collusion.

Failure to reference work that has been obtained from other sources or to copy the words and/or code of another student is plagiarism and if detected, this will be reported to the School's Discipline Committee. If a student is found guilty of plagiarism, the penalty could involve voiding the course. Students must never give hard or soft copies of their coursework reports or code to another student. Students must always refuse any request from another student for a copy of their report and/or code. Sharing a coursework report and/or code with another student is collusion, and if detected, this will be reported to the School's Discipline Committee. If found guilty of collusion, the penalty could involve voiding the course.

1.10 Feedback.

We will endeavour to provide feedback (both verbally in class and in writing on Canvas) within 15 days of submission. Please approach the lecturers for individual feedback on particular questions.

2 Assessment Criteria

Broadly, the assessment criteria are as follows:

70% or higher (A grade)

Structure

- structures assignment effectively to facilitate development of argument.

Content

- displays extensive, detailed and secure knowledge and understanding of the subject
- applies mathematical methods fully accurately to support and develop argument
- demonstrates clear knowledge and understanding of qualitative and quantitative aspects of the question.

Argument

- engages directly with the question and appreciates wider implications and context
- presents a clear, coherent and persuasive argument based on correct mathematics
- displays independence or originality of judgement.

Expression

- uses fluent and accurate prose
- very good standard of presentation.

60-69% (B grade)

Structure

- structures assignment to facilitate development of argument.

Content

- displays extensive and secure knowledge and understanding of the subject
- applies mathematical methods mostly accurately to support and develop argument
- demonstrates sound understanding and knowledge of qualitative and quantitative aspects of the question.

Argument

- engages critically with the question and displays appreciation of the wider implica- tions and context
- presents and develops ideas logically and persuasively
- demonstrates some independence of judgement and initiative.

Expression

- uses clear and generally accurate prose
- good standard of presentation.

50-59% (C grade)

Structure

- broadly structures assignment but organisation of ideas and evidence is sometimes determined by material rather than by the need to develop and support a logical argument.

Content

- displays sound and largely accurate knowledge and understanding of subject
- applies mathematical methods to support and develop argument but there are issues concerning the accuracy of results and the applicability of ideas
- demonstrates limited understanding and knowledge of qualitative and quantitative aspects of the question.

Argument

- displays understanding of the questions set but may lack sustained focus and ap- preciation of the wider context
- states ideas but may not develop them sufficiently or order them in a logical se- quence.

Expression

- prose conveys meaning but lacks sophistication needed to present ideas persuasively
- expression may be clumsy with narrow vocabulary and spelling or grammar errors
- adequate standard of presentation.

Before you start answering the questions you should do the following:

i. Use the command `rm(list=ls())` in programming language R to remove all objects from the R memory.

ii. Load the following packages:
```
library(tidyverse)
library(keras)
library(reticulate)
library(ggplot2)
library(cluster)
library(ClusterR)
library(readxl)
library(magrittr)
library(dplyr)
```

iii. Use the function `getwd()` to locate your working directory. Then, download all the data sets for Parts 2,3 and 4 from **Canvas** and place them into your working directory.

Part 1: Data Science Principles in Insurance [15 Marks]

By conducting your own research, identify the key principles guiding the use of algorithms in insurance companies' decision-making processes. Additionally, provide a brief overview of the measures insurers can adopt to safeguard personal privacy, promote fairness, and maintain solidarity. Ensure all sources and articles used in the response are cited.

[15 marks]

Part 2: Data Manipulation and Wrangling [25 Marks]

For this question, you are given a motor insurance data set called `motor_insurance` with the following columns:

- `caseId`: a numeric column used for identification.
- `attorney`: whether the claimant is represented by an attorney (1 for no and 2 for yes).
- `claimantGender`: the claimant's gender (1 for male and 2 for female).
- `accident_type`: the type of motor accident (0 for minor, 1 for severe, 2 if not applicable).
- `claimantAge`: the age of the claimant.
- `loss`: the total loss incurred to the claimant in thousands.

(a) Display the structure of the data set `motor_insurance` and compute summary statistics for each of its columns.

[**1 mark**]

(b) Calculate the range, interquartile range, 0.5 and 99.5 percentile of the `loss` column. Are there any obvious outliers? Replace these entries with 'NA'.

[**3 marks**]

(c) Delete the entries related to outlying `loss` values above. In insurance, transactional data might appear on a data set as negative claim amount. Whilst there are uses for this, we do not need this right now. Remove the rows with negative values in `loss`. [**7 marks**]

(d) Perform a boxcox transformation on `loss`, record this on a separate column.

[**4 marks**]

(e) Compute the mean claim values amongst men and women. [**4 marks**]

(f) Create a categorical variable for the age of the claimants, with the following groups: (strictly) under 25, from 26 to 35, from 36-42, from 43-72, and 72 or above. [**6 marks**]

Part 3: Principal Component Analysis (PCA) and Linear Regression [30 Marks]

You are given a data set called `marine`. It contains the following variables:

- **vesselVal**: vessel value.
- **claimCount**: number of claims.
- **claimCost**: claim amount.
- **vesselAge**: age of the vessel.
- **distance**: average sailing distance.
- **duration**: the number of policy years.

(a) What two variables are most correlated? **[3 marks]**

(b) Perform a PCA using the `princomp()` function in R and select the least number of principal components that retain 80% of variation in the data.

[10 marks]

(c) Produce a PCA plot. **[3 marks]**

(d) Fit a linear regression model on the principle components selected above. Don't worry about model selection and/or residual analysis. **[6 marks]**

(e) A new policy is written with the following specification:

variable	value
`vesselVal`	22
`vesselAge`	14.5
`distance`	25
`duration`	10

Use the regression model you have fitted above to predict the `claimCost` of this policy.

[8 marks]

Part 4: Deep Neural Networks [30 Marks]

Load the data called `freMTPL2freq` which is given to you. Response variable in this data set is the number of claims and contains several explanatory variables.

(a) Display the structure of the data set and use a Min-Max scaler to normalize all continuous predictors between -1 and 1 and embed the categorical feature components into embedding layers. [**3 marks**]

(b) Construct a 3-layer network with neurons $(q1, q2, q3) = (27, 22, 17)$ and employ the hyperbolic tangent activation function. Additionally, incorporate a non-trainable offset, log(Exposure). Proceed by utilizing the Poisson deviance loss as the objective function, and outline the Nadam optimizer. Subsequently, initiate the training of the network.

[**15 marks**]

(c) Explain the training process for a network with a structure resembling the one mentioned earlier, utilizing the deviance loss of the Negative Binomial distribution with parameters $\mu > 0$ as the mean and $\sigma = 0.8$ as the dispersion.

Provide all the relevant mathematical expressions for training the network. Note in question (c) you are not asked to implement any algorithm in language R.

[**12 marks**]

HINTS

Guidance on Report Structure

You should address the questions posed to you in the list of questions provided as written up solutions in a report structure, where you make sure that you investigate the components of each question according to the amount of effort reflected by the marks allocated.

You introduction should contain a brief overview and brief literature review for Part 1. For Parts 2, 3 and 4, all models used should have formal definition mathematically and all model assumptions or data transformation processes considered should be explained and justified.

All plots and tables should be clearly labeled and have suitable captions and adequate explanation to make then interpretable.

1 page executive summary

Introduction

Data description and preliminary analysis

Methods or models

Results and discussion

Conclusions

Bibliography

Appendix

The appendix should have a table of contents for each R function. All code in R should be provided in the appendix. Full commenting of code should be provided to describe your results. The code should run and be clearly useable – it will also be examined

Requirements regarding R code commenting

For each R function created in this project provide explanatory comments at the front of the script

- Name of the author of the R script/function
- date of writing script
- dependencies on any other packages etc in R
- version of R and package versions required for the script
- what you aim to achieve with the script e.g. a brief description

Furthermore, all R code should be commented with sufficient explanation and if you used other people's functions or built in R functions you should explain their purpose and use (i.e. settings) and your choices for these should be justified **briefly**.

References to any functions not written by you should also be provided in the report. e.g. cite the vignette from the R function and its source or url in your references section of your report.

For each piece of code you create you should clearly explain the input and output and provide in your report the results obtained from running the code; and any relevant explanations or plots requested to interpret the output of the code.

- All code written for the project should be included in the Appendix of the report and it should be clearly labeled with regard to the question it addresses and its purpose.

Department of Actuarial Mathematics and Statistics

Full name:

Matriculation number:

Degree:

F71RA Machine Learning for Risk and Insurance: Project

Plagiarism declaration:

I confirm that I have read and understood: (a) the note on Plagiarism and collusion in the assignment handout; (b) the Heriot-Watt University regulations concerning plagiarism.
I confirm that the submitted work is my own and is in my own words.
I confirm that any source (aside from course notes and lecture material) from which I obtained information to complete this assignment is listed in the assignment. Any sources not listed in the assignment are listed here:

Apart from the lecturer, I discussed the assignment and shared ideas with the following people:

Signature
Date
(Please attach to the front of your completed assignment for submission.)

[END OF PAPER]