Nov 2024 ST3189 CA1

Instructions

Paper Release: Wed 6 Nov 2024.

Submission Deadline: Thu 14 Nov 2024, 6pm. Submission via Canvas.

1. The estimated duration of this assignment is **2 hrs**. The actual hours will depend on your familiarity of the concepts and coding.

2. Type all your answers (including software output screenshots and code where necessary) into a Microsoft Word Document. All answers must be fully contained inside the word document. i.e. do not ask examiner to refer to any other sources to see your answers. Save and rename the submission file with your registered full name. There is only one file to be submitted.

3. This is an open-book individual homework assignment, and you are allowed to use internet for research.

4. If you use someone else's work or opinion, quote and cite properly.

5. You are allowed to use the following software to answer the questions:

   - R and RStudio

   - Python

   - Microsoft Excel and Microsoft Word

6. Your submission will be graded on correctness, compliance to instructions and presentation. Label your answers in the word document to correspond to the question number/part. Many questions are designed to be open-ended and have several possible answers or ways to express your answer. Strive to be concise and precise in your answers. Excessive words or quotes will be penalized. You are not expected to write essay, and excessive words signal that you do not know what is important vs unimportant.

7. You may use more than one page per question, if necessary, but start each answer to a new question number on a new page.

8. This CA is designed to help your score in the Coursework project so that you can do well in the overall official grade.

9. You may submit up to maximum 3 times before the deadline into Canvas. Only the latest submission will be graded.

10. Non-submission or poor scores may affect your eligibility to sit for official exam. Submission is only accepted via Canvas. Late submissions will not be marked.

CA1 Question Paper

ML Generated Insurance Premium

Introduction

An insurance company wants to use Machine Learning to generate insurance premium payable and propose the insurance offer to the client. This will shorten the sales process and eliminate effort and time to process the request. Currently, the sales cycle relies on insurance agent, actuaries, admin staff, and take too much time and effort.

To evaluate the feasibility, a sample of hospitalization insurance premium and client profile data (**premium2.csv**) was handled to you. A data dictionary is provided in Appendix A.

1. Create the BMI variable based on CDC definition[1]. Show your code or formula used, and the BMI values for the first three rows of data provided.

2. Explore the data and report 3 key findings.

3. Using Linear Regression:
    a. Explain how you will select the “optimal” subset of X variables in your final linear regression model.
    b. Do a 70-30 train-test split and report on the testset RMSE. Explain the meaning of RMSE.
    c. Is BMI or Gender important in determining premium?

4. Explain how the above analysis can be improved in less than 500 words.

5. Show your Rcode or Python code as an Appendix in your word document, clearly labelling the section of the code used for each question. [Note that the code in Appendix is for examiner’s reference and will not be marked.]

[1] https://www.cdc.gov/healthyweight/assessing/index.html

Appendix A

Age:	Age of the client (years).
Diabetes:	Presence (1) or Absence (0) of the disease.
HighBloodPressure:	Presence (1) or Absence (0) of the condition.
Transplant:	Organ Transplant Recipient (1: Yes, 0: No).
ChronicDisease:	Other Chronic Disease besides Diabetes or High Blood Pressure (1: Yes, 0: No).
Height:	Height (cm).
Weight:	Weight (kg).
Allergy:	Known Allergy (1: Yes, 0: No).
CancerInFamily:	Does any family member had/have Cancer (1: Yes, 0: No).
NumMajorSurgeries:	Number of Major Surgeries Done, excluding Organ Transplant.
Gender:	(1: Male, 0: Female).
Premium:	Annual Premium Payable by client (Singapore Dollars)