ME 781: Engineering Data Mining and Applications

Course Project

Note:

1. **This project work carries 20% weightage**
2. It is a group project with maximum 3 participants (it is mandatory to register your group using the link provided on Moodle, without which submissions will not be evaluated)
3. Based on your group number, only process the appropriate data file. Eg: Group number 4 should process data file **g-4.csv**. **No credit will be given if wrong data file is processed**.
4. All submissions will have to be to Moodle. There can be one submission per group.
5. Last date for submission: Friday, April 19, 2019 – 2355 Hrs.
6. **Unfair means such as copying of code / report / data set will attract 100% penalty.**

Submission Guidelines

1. All the artifacts mentioned in the points below (documents / code / data) should be zipped into one ZIP file and uploaded to Moodle. It should be possible to extract the ZIP file and read your report / run Python programs without additional moving around of programs / data. Include a README.TXT file to explain the contents of your folder(s).
2. Create a neat final report document (PDF) that contains all analysis, conclusions, assumptions, and justifications, with charts, images, properly numbered and explained. In this report you may refer to Python code files and include small and relevant code snippets as required for clear explanation, but do not dump the entire Python code or output generated by Python Code.
3. All Python code files should be appropriately named and well documented. All Python code should be executable without errors in Python 3.5 or above. All Python code files should be submitted – **preferably as Jupyter Notebook(s)**.
4. Your work requires the generation of additional data files, submit those files too.

Project Description

Part-1 (10 Marks)

Using the data file assigned to you, create a working file, as follows, to simulate **missing data**.

- 10% data from **each** predictor should be independently and randomly knocked off and substituted with **NA** (this will ensure that an entire row **does not** consist of NAs)
- Name the working file as g-XX-w.csv and it should be part of your submission.

Using the working file analyse the data and generate your best possible ML model. In your report explain your overall strategy and all steps, with reasons, taken for generating the best model. Justify with evidence why your final proposed model is the ***best model***.

Part-2 (10 Marks)

Using the working file g-XX-w.csv created in Part-1, do the following:

1. Discretize the response variable Y into a qualitative variable with 5 levels
    - Hint: Find out the range of Y, divide it into 5 equal intervals 1…5, and replace the actual values of Y with the corresponding interval number.
2. Likewise discretize any two of the predictor variables into qualitative variables with 3 and 5 levels respectively.
3. Name the modified data file as g-XX-w2.csv, and it should be part of your submission.

Using this working file analyse the data and generate your best possible ML model. In your report explain your overall strategy and all steps, with reasons, taken for generating the best model. Justify with evidence why your final proposed model is the ***best model***.

Evaluation Criteria:

1. Completeness and correctness of approach (ML model creation process and steps): 60%
2. Quality of report / documentation: 30%
3. Code quality : 10%

Important: Unfair means such as copying of code / report / data set will attract 100% penalty.

ooo000ooo