ANL252

End-of-Course Assessment – July Semester 2021

Python for Data Analytics

_____

INSTRUCTIONS TO STUDENTS:

1. This End-of-Course Assessment paper comprises **SIX (6)** pages (including the cover page).

2. You are to include the following particulars in your submission: Course Code, Title of the ECA, SUSS PI No., Your Name, and Submission Date.

3. Late submission will be subjected to the marks deduction scheme. Please refer to the Student Handbook for details.

IMPORTANT NOTE

ECA Submission Deadline: 17 September 2021, 12 noon

ECA Submission Guidelines

Please follow the submission instructions stated below:

A - What Must Be Submitted

You are required to submit the following item for marking and grading:

- *A Report*

Please verify your submissions after you have submitted the above item.

B - Submission Deadline

- *The report is to be submitted **by 12 noon** on the submission deadline.*

- *You are allowed multiple submissions till the cut-off date for the report.*

- *Late submission of the report **will be subjected to mark-deduction scheme** by the University. Please refer to Section 5.2 Para 2.4 of the Student Handbook.*

C - How the report Should Be Submitted

- *The Report: submit online to Canvas via TurnItIn (for plagiarism detection)*
  - *please ensure that your Microsoft Word document is generated by Microsoft Word 2007 or higher.*
  - *the report must be saved in .docx format.*

- *Avoid using a public WiFi connection for submitting large video files. If you are using public wireless (WiFi) connection (e.g. SG Wireless at public areas), you might encounter a break in the connection when sending large files.*

D – Please be Aware of the Following:

Submission in hardcopy or any other means not given in the above guidelines will not be accepted. You do not need to submit any other forms or cover sheets (e.g. form ET3) with your ECA.

*You are reminded that electronic transmission is not immediate. The network traffic may be particularly heavy on the date of submission deadline and connections to the system cannot be guaranteed. Hence, you are advised to submit your work early. **Canvas will allow you to submit your work late but your work will be subjected to the mark-deduction scheme**. You should therefore not jeopardise your course result by submitting your ECA at the last minute.*

It is your responsibility to check and ensure that your files are successfully submitted to Canvas.

E - Plagiarism and Collusion

Plagiarism and collusion are forms of cheating and are not acceptable in any form in a student's work, including this ECA. Plagiarism and collusion are taking work done by others or work done together with others respectively and passing it off as your own. You can avoid plagiarism by giving appropriate references when you use other people's ideas, words or pictures (including diagrams). Refer to the APA Manual if you need reminding about quoting and referencing. You can avoid collusion by ensuring that your submission is based on your own individual effort.

The electronic submission of your ECA will be screened by plagiarism detection software. For more information about plagiarism and collusion, you should refer to the Student Handbook (Section 5.2.1.3). You are reminded that SUSS takes a tough stance against plagiarism or collusion. Serious cases will normally result in the student being referred to SUSS's Student Disciplinary Group. For other cases, significant mark penalties or expulsion from the course will be imposed.

The text file "ship.csv" contains the data of reported ship damage incidents. They are taken from P. McCullagh and J.A. Nelder and can be found in their book *Generalized Linear Models* (New York: Chapman & Hall, 1983). The study's main objective is to investigate how the number of ship damage incidents (variable: Y) is related to the following independent variables:

- the aggregate months of service (MS),
- ship type (T: 1−5),
- year of construction (A: 1 for 1960-64, 2 for 1965-1969, 3 for 1970-1974, and 4 for 1975-1979),
- period of operation (P: 1 for 1960-1974, and 2 for 1975-1979).

Use JupyterLab to solve the following questions and attach screenshots to demonstrate the output of your Python program to each task.

Question 1

Prepare the ship data stored in "ship.csv" for future analytics tasks using functions and methods of NumPy and pandas, respectively.

(a) Design your own Python program to carry out the following tasks:

(i) Read in "ship.csv" as pandas DataFrame called "ship". Since there are 6 observations where MS and Y are "." to indicate that they are missing values, declare this character as missing values in your program accordingly.

(ii) Since the variable names of this dataset are rather short and do not really describe the nature of the variables, rename the ship types to "types", construction years to "c_years", operation periods to "o_periods", the aggregated months of service to "s_months", and the number of incidents to "incidents".

(iii) For better understanding of the data, compute the average service months and the average number of incidents for the cross-products of every category in types and operation periods. The averages should be rounded to the nearest integers. Store the resulting table to an object named "shipgroup".

(iv) Replace the missing values in the variable "s_months" and "incidents" by the respective means of the other ships that share the same type AND the same operation period. Add comments to elaborate your Python program as well.

(v) Construct a Python program to save the target variable "incidents" in a pandas DataFrame named "Y".

(33 marks)

(b) Except for the months of service and number of incidents, all the other variables, including "types", "c_years", and "o_periods" are actually nominal and not interval/ratio.

(i) Perform an appropriate data type conversion for these variables so that they can be recognised as categorical variables.

(ii) Construct Python code to convert all categorical variables to dummy variables and save the result as a pandas DataFrame named "X".

(iii) Researchers suggest that the aggregated months of service of each ship must be scaled down due to its wide range of values. Perform a log-transformation of this variable in the DataFrame and name the transformed variable "log_s_months". The transformed variable should be attached to both DataFrames "X" and "ship".

(14 marks)

(c) Normally, we shall split the DataFrame into training and testing datasets to evaluate the predictive power of the model. Study the dataset carefully and explain why it is not sensible to split the DataFrame here, and we shall use the entire dataset for training purpose instead.

(8 marks)

(d) We shall now save the prepared DataFrame "ship" as a new csv text file called "ship_prepared.csv". Furthermore, we shall also create a database called "ship.db" and export the DataFrame to the database as tables. Write a Python program to carry out these two tasks.

(15 marks)

Question 2

In their book *Generalized Linear Models* (New York: Chapman & Hall, 1983), the authors P. McCullagh and J.A. Nelder used the Poisson regression to study the ship dataset. Poisson regression is a special case of the generalised linear models in which the target variable, or dependent variable, is Poisson distributed. Since one of the main application areas of Poisson regression is to fit linear models on count data, we can therefore use Poisson regression to predict the number of incidents (which are also counts) given some input variables.

Mathematically, Poisson regression is a linear model in which the expected value of the target variable Y is calculated by

$$\log \mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k,$$

where β_0 is the intercept, β_1, β_2, ... , β_k are the coefficients of the independent variables X_1, X_2, ..., X_k. E(Y) is the predicted, or expected value of Y, which will be transformed by the natural logarithm function.

(a) Find the corresponding scikit-learn module in the official website of scikit-learn and discuss the corresponding module, estimator, fit and predict functions, as well as their parameters in your own words.

(10 marks)

(b) Analyse the data by fitting a Poisson regression based on the DataFrames X and Y generated in Question 1. Follow the instruction in the official website and report the parameters of the estimated model. Create a Python program to fit a Poisson regression and generate a table or a DataFrame to present the coefficients with the corresponding labels.

(10 marks)

(c) The deviance of Y and its expected value E(Y), estimated by the model constructed in c), measures the goodness of fit of the model. The lower the deviance, the better is the model. Below is the equation of how it should be calculated.

$$D = 2 \sum_{i=1}^{n} \left\{ Y \log \left[\frac{Y}{\exp(\mathbb{E}(Y))} \right] - [Y - \exp(\mathbb{E}(Y))] \right\}$$

If Y = 0, the expression log[Y/exp(E(Y))] will be taken as zero. Employ your own Python program to compute D <u>without</u> using the score() function of the scikit-learn package.

(10 marks)

----- END OF ECA PAPER -----