DS203-2023-sem2: Exercise – 1

Submissions due by: Jan 21, 2024, 11:55pm

Part - A

1. Review the closed form derivations for the SLR coefficients 'a' and 'b' provided in the uploaded file **Simple Linear Regression Derivation.pdf**
2. Review the sample data set provided in the uploaded file **data-set-for-SLR-2024-01-12.xlsx** and the calculations of **ycap, e, etc.** in the sheet **data-set-for-SLR-Class**.
3. Review the error metrics and their calculations.
4. Install and enable the **Data Analysis Toolpak** in MS Excel (Google how to do it ..)
    - If you are using macOS and do not have access to Excel, download and install LibreOffice and use the Excel equivalent package **Libreoffice Calc**. Functionality equivalent to the Data Analysis Toolpak is available through '**Data / Statistics**' or '**Data / >> / Statistics**' menu options.
5. Use the inbuilt Excel / LibreOffice Calc **Linear Regression** / **Regression** functionality to create an SLR model – as illustrated in the sheet **SLR-Excel**.

Part – B

In this part of the exercise, you will work on fresh data to create a Simple Regression Model.

6. Load the data file **e1.xlsx**. Note that this file contains two tabs **train_data** and **test_data**. The total data available for analysis is randomly divided into these two sets usually using an 80:20 split. This allows the ML model to be created using the **train_data**, and tested for the model's general applicability (ie. on unseen data) using the **test_data**.
7. For the train_data, visualize **y v/s x** using a scatter plot
8. Create SLR models using the train_data. Use the following methods to create the SLR models:
    a) Using the closed form equations for 'a' and 'b'
    b) Using the in-build Linear Regression / Regression functionality of Excel / LibreOffice Calc
9. Calculate the following using the train_DATA: **ycap, e, e_sq, MAE, SSE, MSE, RMSE**. Analyze and record your observations.
10. Superimpose **ycap v/s x** on the scatter plot created in step 7. Analyze and record your observations.
11. Create a scatter plot **e v/s x** and analyze and comment on its characteristics.
12. Create a **histogram of e** and comment on its characteristics.
13. Using the coefficients calculated in steps 8(a) / 8(b), repeat steps 9 – 12 using the **test_data**
14. Compare the error metrics and error plots for **train_data** and **test_data**. What do they indicate?
15. Create a document by neatly capturing all the above plots and your analyses and comments.
16. Convert the document into a PDF. Name of the PDF should be **E1-your-roll-number.pdf**. Upload it to the assignment submission point E1.

oooOOOooo