Exercise – 2: DS203-2023-sem2

Submissions due by: Jan 31, 2024, 11:55pm

This exercise is aimed at understanding the relationship between data quality, sample size and the various metrics obtained after performing Linear Regression.

Part - A

1. Three datasets E2-set1.csv, E2-set2.csv, E2-set3.csv have been added to Moodle. You are required to pre-process these datasets to gain a good understanding of their characteristics. Ensure that you comment on the quality of each dataset in comparison with the characteristics of the other datasets. Create an analysis and include it in your report.
    - Questions to be answered: Based on what calculations will you understand the characteristics of the dataset? What charts / plots will you create? How will you determine the quality of the dataset - what calculations will you do and which statistics will you use?
2. A Notebook E2-process-data.ipynb has also been added to Moodle. Review and understand the flow of the Python program and the outputs generated by it. Include a summary in your report.

Note:

- You will notice that the Notebook generates sufficient data for each sample size. The following part of this exercise requires you to understand the trend of the values i) across the samples for a given sample size ii) across the sample sizes for a given dataset iii) across the datasets for the same sample size! **So, carefully understand the tasks and plan your data analysis!**

Part - B

3. Process each of the datasets (using the above Notebook) for the following **sample sizes {5, 10, 20, 50, 100}** and understand the following:
    - For each of the datasets, and across the datasets, how does the sample size impact the calculated statistics (ie. the coefficients, their p-values, their 95% Confidence Intervals) and their metrics (ie. R2, MSE, F-Statistic and its p-value). How can you explain the observed impact? Please do not only state the obvious! Enough data is generated for each sample size, therefore your critical analysis and reasoned explanation – based on theory - are expected. Design and create the required Tables / graphs / charts to summarize your observations and provide your detailed explanation. Include all this in your report.
4. Consider E2-set1.csv. Explain why the average R2 value (average across samples for a given sample size), does not improve even when the sample size is increased from 10 to 100. What happens to the average values of the other metrics? What conclusion will you make about the usability of the models based on this dataset? Will you use the model(s) at all? Which ones? Why?
5. Consider E2-set3.csv. Explain why the average R2 value, across samples, is high even for the very small sample size of 5. What happens to the average values of the other metrics ? What conclusion will you make based on these observations?
6. Based on all the above, create your own guidelines to help you assess the quality and acceptability of an LR model.
7. Create a document by neatly capturing all the above analyses and comments.
8. Convert the document into a PDF. Name of the PDF should be **E2-your-roll-number.pdf**. Upload it to the assignment submission point E2.

Note: You are free to create your own additional experiments and report them!

oooOOOooo