CS Rubric – Case Study Reproduction

Course: DS 4002
Due: by the end of the semester, via Canvas
Submission format: Upload link to GitHub repository to Canvas
Individual Assignment

Why am I doing this? This is your opportunity to synthesize the lessons learned during this course and complete a project based on a case study file. The deliverable you will create is a reproduction of a past project, to test your technical abilities in a real-world scenario.

- Course Learning Objective: recreate a past project to practice your technical skills

What am I going to do? You will research the case study provided and become familiar with the motivation for the project and the necessary technical elements required. Then, you will use your knowledge to recreate the case study, including code and output graphs. Deliverables include a GitHub repository and data set – to provide resources like code and data. All of this will be submitted electronically via a link to a GitHub repository built for the case.

How will I know I have Succeeded? You will meet expectations on CS Case Study Reproduction when you follow the criteria in the rubric below.

Spec Category	Spec Details

Formatting	Submit the following to Canvas to earn credit:
	• A link to a GitHub repository that contains the following information:
	  ○ A folder containing all data used to complete the case study
	  ○ A folder containing all scripts used to complete the case study
	  ○ A folder containing all outputs generated from running the code
	  ○ A README explaining the case study and the contents of each folder

Data folder	• Goal: have all the data necessary to complete the case study stored in one location • Include any data that was initially provided in the case study • Include any data that was generated by the code (whether it was pre-existing code from the case study or code that you wrote yourself) ○ There should, at minimum, be the original data file and a cleaned data file containing the data that was used to train the model

	○ Both of these files can be found on the GitHub link provided ● Ensure that the titles of the files are descriptive, and that the commits explain what each file is
Scripts folder	● Goal: have all the scripts necessary to complete the case study stored in one location, and effectively commented so anyone could use them ● Include all scripts used in the case study (including exploratory data analysis, data cleaning, and the actual model code) ○ There should, at minimum, be code for some type of regression model that successfully predicts air pollutant concentrations for a future month ○ Code for a Random Forest regression model can be found on the GitHub link provided ● Ensure that all code is commented clearly, so that anyone could look at the code and understand what it does and how it works ● Ensure that the titles of the files are descriptive, and that the commits explain what each file is
Outputs folder	● Goal: have all the outputs generated by the code stored in one location ● Include any graphs or tables generated by the code ○ There should, at minimum, be two graphs for each of the five pollutants: one graph for the validation step, showing the predicted concentration of the pollutant against the actual concentration of the pollutant for a past (known) month; and one graph for the prediction step, showing the predicted concentration of the pollutant for a future (unknown) month ○ Additionally, include some type of metric demonstrating how accurate the model is; for example, a table of root mean squared errors or R^2 values for the various graphs. How you display this information is up to you! ● Ensure that the titles of the files are descriptive, and that the commits explain what each file is
README file	● Goal: a descriptive file that explains the contents of the GitHub repository, and how you used each component of the folders to produce the required result ● Clearly outline the contents of each folder

	● Write out your process for completing the case study. Be sure to include: ○ What steps you took to reproduce the case study ○ The order you completed these steps in ○ What data and script files you used ○ What output graphs/tables you produced ● Write a brief review of the case study. Be sure to include: ○ What you liked/disliked about the case study ○ What you thought the author did well ○ What you thought the author should adjust to make the case study better (more clear, more engaging, etc)

Acknowledgements: This structure is pulled from Streifer & Palmer (2020).