Cloud for AI

Assignment Task 3.

Task 3: Design your deployment solution.

For this final part of the project, you will plan out the deployment of your machine learning model (without actually deploying it). The aim is to get familiar with the essential aspects of deploying a model in a production environment, including practical considerations around accessibility, cost, scalability, and performance.

Objectives of the assignment

Part 1:

- Build a GitHub Repo for your application
- Design an endpoint using Flask/FastAPI etc
- Create a dockerfile

Part 2:

- Write a deployment report

Part 3 (Bonus)

- Deploy the model using Vercel (or any other platform)

We have talked about Deployment technologies
Follow these guidelines. Keep it simple.

For Part 2, briefly answer the following questions:

1. End-User Access:

- Describe how the model is made available to the end user. Think about the user experience:
  - How will the users interact with the app?
  - How will users access your model's predictions?
- Consider ease of use, response time, and the expected user workflow.

2. Balancing Latency and Costs:

- Discuss how you would balance the response time (latency) of your model with the associated costs.
  - Describe how you would optimize costs without sacrificing necessary performance, especially if the model is used in real-time.

3. Deployment Location:

- Specify where the model will be deployed:

    - **Cloud**
    - **On-premises**:
    - **Edge devices**

4. Scaling the Model:

- Briefly describe your approach to scaling.
- Consider potential future growth. How will your deployment plan handle an increase in users or requests?

5. Inference Mode:

- Determine whether your model will use **batch inference** or **online (real-time) inference**:
- Justify your choice based on the needs of your project.

Remember there are no perfect answers. Your explanations are what will be graded!

You will be evaluated on the following:

- Clarity of Explanation
- Comprehensiveness of your plan: Make sure your plan covers all essential aspects of deployment. You should have a complete solution.
- Tradeoff consideration
- Justification of your choices.

To deliver:

- A link to your GitHub repo
- A 2-3 page report.

Part 3:

This part is optional and will serve as a bonus if completed.

Use the code examples provided.

Attention: In case you use AWS or Google Drive, pay attention to your expenses.

Final Notes:

If you have any questions, concerns, or anything you want to discuss please email me so we can sort it out!

On Appendix please provide clear descriptions on the AI tools used, why and how.