\*\* Rubric Questions & Answers \*\*

Dr. Shaista Hussain

MLE 10

FOURTHBRAIN

\

\* QUESTION 1:

>> What is ML model deployment?

Model serving deployment is the process of deploying a trained machine learning model in a production environment. This allows the model to be used in real-world applications, accessible by users who need to use the model's predictions. An ML model is deployed using one of may available platforms such as AWS SageMaker, Google Cloud ML Engine, Microsoft Azure, or IBM Watson Machine Learning. The most commonly used method today is model serving or model as a service.

\* QUESTION 2:

>> What are 3 advantages of deploying using Model Serving methods Vs. deploying on GitHub Pages or HuggingFace for free?

This architecture effectively separates the application from the model through an API, simplifying organizational processes, model versioning and reuse, seamless updates with phased rollouts and hardware separation allowing the application and model server to have diverse hardware (CPU, GPU, FPGA) and independent scaling upon requests.

GitHub Pages is a static web hosting service, which allows you to deploy HTML, CSS, JavaScript, and other static files to the web. You can use GitHub Pages to host a website for your project or to host a static site for your model.

HuggingFace is a platform for hosting and sharing pre-trained machine learning models. On HuggingFace, you can store and deploy your models as well as access other pre-trained models. Additionally, HuggingFace provides a platform for sharing and collaborating on models, making it easy to find and use the right models for your project.

\* QUESTION 3:

>> What is Causal Inference and How Does It Work?

Causal Inference is the process of inferring cause and effect relationships from data. It is a form of data analysis that helps scientists, researchers, and analysts to determine the causes of certain outcomes and events. It is a key tool for understanding the underlying structure of a system, as well as for developing predictive models and forecasting future outcomes.

To understand how causal inference works, it is important to consider the components of a causal system. In a causal system, each component is viewed as a cause that can lead to an effect. A causal system is composed of two types of variables: independent variables (or causes) and dependent variables (or effects). The goal of causal inference is to identify the causal relationships between these two types of variables.

To do this, causal inference uses various statistical techniques, such as regression analysis and path analysis, to investigate the relationships between the independent and dependent variables. The results of these analyses can then be used to understand the underlying structure of a system, as well as to develop predictive models and forecast future outcomes.

\* QUESTION 4

>> What is serverless deployment and how its compared with deployment on server?

Serverless deployment is a cloud computing model in which applications are deployed without the need for servers or other infrastructure. Instead, the applications are run on a series of virtual machines or containers, which are managed by the cloud provider. This allows for more efficient scaling and easier management of the application.

Compared to deployment on a server, serverless deployment does not require the user to manage the underlying infrastructure or servers. This makes it easier for developers to focus on building the application itself, without having to worry about server maintenance or scalability. Additionally, serverless deployment allows for more efficient scaling and automated updates, which can help reduce the cost of running the application.

REFERENCES

1. What is serverless? https://www.ibm.com/topics/serverless

2. Serverless Deployment: How It Works and Practical Examples https://www.sentinelone.com/blog/serverless-deployment/

3. Serverless computing vs. containers | How to choose | Cloudflare https://www.cloudflare.com/learning/serverless/serverless-vs-containers/

4. Deploying to AWS - Serverless Framework https://www.serverless.com/framework/docs/providers/aws/guide/

5. Causal Inference: Trying to Understand the Question of Why | by ...

https://towardsdatascience.com/implementing-causal-inference-a-key-step-towards-agi-de2cde8ea599

6. What is Causal Inference and How Does It Work? - Manning https://freecontent.manning.com/what-is-causal-inference-and-how-does-it-work/

7. Causal Inference - an overview | ScienceDirect Topics https://www.sciencedirect.com/topics/social-sciences/causal-infe

8. A guide to ML model serving https://ubuntu.com/blog/guide-to-ml-model-serving

9. ELS-RD/transformer-deploy: Efficient, scalable and ... - GitHub https://github.com/ELS-RD/transformer-deploy

10. MLOps: An overview to Machine Learning Demos | by Li Yin | Medium https://liyin2015.medium.com/how-to-quickly-do-machine-learning-demo-52c49cc06e92?source=user\_profile---------0----------------------------

11. huggingface GitHub Topics GitHub https://github.com/topics/huggingface