Streamlining ML model Deployment

Team2 - Ankitha Shetty, Kirtana Kirtivasan, Swetha Vijaya Raju, Tejaswini Kambhampati

About The Company (Fictitious)

- Lead is a GPS-enabled navigation application that uses real-time user location data and user-submitted reports to suggest optimized routes.
- Company size : 100-150 employees

Business Problem

- Lead's Data science teams generate tens of ml and dl models, (using python) from different verticals, on a daily basis which needs to be translated to job schedulers to deploy them into production environment
- Lead's Engineering and Operation teams are also required to monitor the health of these data pipelines
- This process is time consuming,extremely mechanical and iterative in nature
- They want to find the appropriate service/solution provider to standardize, streamline their model deployment and monitoring needs

Project Objectives

Technical Objectives:

- Avoid latency and speed up the deployment process
- Easy to set up service
- Simplified logging
- Achieve version management
- Initiate batch requests to the same model, so hardware is used efficiently.

Project Objectives

Political Objectives

- Promotes efficient collaboration among ML Engineers, Data Scientists, Product Managers and Site Reliability Engineers
- Foster compatibility with the existing system

Financial Objectives

- Free to use open source platforms thus reducing the deployment cost significantly
- Modular cost, based on the feature you use

Recommendations: AWS SageMaker

AWS Sagemaker is a powerful service provided by Amazon. It gives ML developers the ability to build, train, and deploy machine learning models quickly.

Advantages:

- It simplifies the whole machine learning process by removing some of the complex steps, thus providing highly scalable ML models.
- The cost is modular, based on the feature you use.
- This accelerates model production and deployment with minimal effort and cost.
- Lead is already using AWS ecosystem for data storage.

Recommendations: MLFlow

The platform can be used for ML deployment by individual developers as well as teams. It can be incorporated into any programming ecosystem. The library is built to satisfy various technological needs and can be used with different machine learning libraries.

Advantages:

- An open-source tool
- Easy to set up and user-friendly UI
- The logging is simplified, so it’s easy to run experiments
- When it comes to training, tuning and deploying ML models, it brings transparency and standardization

Project Benefits

- Increase ROI
- Improve customer experience
- Reduce manual labour
- Increase collaboration between the teams involved
- Enhance scalability and reproducibility

References

- https://neptune.ai/blog/best-8-machine-learning-model-deployment-tools
- https://www.element61.be/en/resource/4-ways-how-mlflow-can-facilitate-your-machine-learning-development
- https://www.truefoundry.com/
- https://github.com/truefoundry
- https://nimblebox.ai/
- https://stackoverflow.blog/2020/10/12/how-to-put-machine-learning-models-into-production/
- https://www.datacamp.com/tutorial/tutorial-machine-learning-pipelines-mlops-deployment