MLE-26 Architecture Overview

SAM ROSKA

Folder structure is intentional to separate files as the application could grow to include multiple models or multiple data processing requests. Also, it's just easier to understand and read.

model-app >
- test >
- models>
- data>
- api>

The application is is build using FastApi for the "/predict" endpoint. This library is very simple to implement and automatically includes documentation of endpoints. After launching a local version of the application visiting localhost:/docs#/ allows for sending request to the api without needing additional applications like Postman.

Data validation is handled by Pydantic. After some research I found that it was common practice to use this library in ML models. There are built in methods for validating data and it allows customization. By use of this library there is no need to create a complex data structure used within the application. I tried to use the Pydantic data objects like an OOP class used throughout the application —however I might change that in a future revision and instead convert the object to a Pandas dataframe.

Data processing responsibility is abstracted to the script Segmentation.py. This python script handles data cleaning/formatting required for the segmentation model. All of the operations on a Pandas dataframe are contained within appropriate methods for pre-processing and post-processing of model results. I considered separating the model predict and that data cleaning, however the script was manageable to read so I kept it all together. In a future implementation I'd include ability to retrain model with new data file.

The build portion of the application includes a Dockerfile for containerization. Containers are relatively easy to implement and scale with an orchestration tool like Kubernetes. The "replicas:" tag permits for scaling containers built using the same image. I made assumptions that a cluster already exists and networking protocols already established.

Local build documentation is included in the README.md of the project.