**Project Assignment: Machine Learning Model Deployment**

Assignment for this week was to setup an environment for my ML-Ops projects. Here I will try explaining every part of the process of setting up my ML-Ops environment.

**Find a ml project.**

First, I took at project from another course I’m doing, but it could have been one of my own ml projects. Goal is to make the ML-Ops flow, so it doesn’t really matter much.

**Created a repository**

I created a project on GitHub called [**Practical-MLOps-assignment-3-model-deployment**](https://github.com/markat1/Practical-MLOps-assignment-3-model-deployment) .

**Launch.json**

Hereafter, I added a launch.json that basically explains how I want ML-Ops project should be run within my environment, which in this case is visual studio code. Here’s the configuration.

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse

Overall what you can do in launch.json is actually set up different parameters for the whole project.

**Created an environment for the whole project.**

I then made sure that my ML-Ops project ran in it’s own environment by using

python -m venv venv and then ran source venv/bin/activate

**Main.py**  
As you can see configuration calls a main.py.

Main.py mainly explains what kinda path should be run in a local or cloud environment. Here we also distinguish between training and inference.

So in order to now make training or inference process collide with each other we use Subprocesses, so each run runs it’s own process.

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse

**Train.py**  
So next I added a train and inference py.

Train is where we train the model.

I chose here a project that in the end is used for predicting the amount of time it would take a carrier to deliver food in India.

Training data here is from <https://www.kaggle.com/datasets/gauravmalik26/food-delivery-dataset/data> and I’m using xgboost in order to predict estimation of delivery time.

You can see in train.py the repository for more information.

**Ready for a spin of training locally**

Okay, so now the is ready for it’s next phase, which is to try spin up an see if we get a result back from training.

So I click the  and got the following result, which seems to respond with a response from training the model. That part works!

Et billede, der indeholder tekst, skærmbillede, software, Font/skrifttype

Automatisk genereret beskrivelse

**Creating a docker build and push it up to sagemaker.**

Okay, so now the is ready for it’s next phase, which is to deploy it to sagemaker.

I created a new repository in aws ECR

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

I then made an docker image of my project with docker build and run a script to push it up to ECR repository.

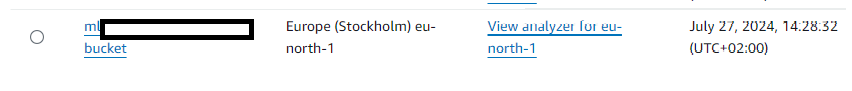
Et billede, der indeholder tekst, skærmbillede, software

Automatisk genereret beskrivelse

My script is in the repository called build\_and\_push.sh, which is called by ./build\_and\_push.

**Create a training job on sagemaker**

Last part was to make project able run my ECR image on amazon as a job, so I started creating an S3 bucket for my training files.



After it was created I added train.csv in S3 bucket.

Et billede, der indeholder tekst, skærmbillede, software, Webside

Automatisk genereret beskrivelse

I was then ready to create a job, which I did running python orchestrator.py

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, software

Automatisk genereret beskrivelse

Then I can this job in aws by looking in amazon sagemaker’s training jobs.

Et billede, der indeholder tekst, Font/skrifttype, logo, hvid

Automatisk genereret beskrivelse

Seems like it completed Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, linje/række

Automatisk genereret beskrivelse

I can check and see if there’s now a output model in s3, which seems to be correct as you can see here. Et billede, der indeholder tekst, skærmbillede, software, nummer/tal

Automatisk genereret beskrivelse