

Tracking. MLflow and load balancing setup

GitHub repository: <https://github.com/xphoenixua/brain-invaders-mlops>

System startup

- `docker compose up --build -d`
- `docker compose -f airflow-compose.yaml up --build -d`

namely,

- `user@MacBook-Pro braininvaders_mlops docker compose ps`

NAME	CREATED	STATUS	IMAGE	PORTS	COMMAND	SERVICE
braininvaders_mlops-airflow-scheduler-1	7 minutes ago	Up 7 minutes (healthy)	apache/airflow:2.10.5	8080/tcp	"/usr/bin/dumb-init ..."	airflow-scheduler
braininvaders_mlops-airflow-triggerer-1	7 minutes ago	Up 7 minutes (healthy)	apache/airflow:2.10.5	8080/tcp	"/usr/bin/dumb-init ..."	airflow-triggerer
braininvaders_mlops-airflow-webserver-1	7 minutes ago	Up 4 minutes (healthy)	apache/airflow:2.10.5	0.0.0.0:8080->8080/tcp	"/usr/bin/dumb-init ..."	airflow-webserver
braininvaders_mlops-airflow-worker-1	7 minutes ago	Up 7 minutes (healthy)	apache/airflow:2.10.5	8080/tcp	"/usr/bin/dumb-init ..."	airflow-worker
braininvaders_mlops-model-serving-challenger-1	8 minutes ago	Up 2 minutes (healthy)	braininvaders_mlops-model-serving-challenger	8000/tcp	"uvicorn main:app --..."	model-serving-challenger
braininvaders_mlops-model-serving-champion-1	8 minutes ago	Up 7 minutes (healthy)	braininvaders_mlops-model-serving-champion	8000/tcp	"uvicorn main:app --..."	model-serving-champion
braininvaders_mlops-model-serving-champion-2	8 minutes ago	Up 7 minutes (healthy)	braininvaders_mlops-model-serving-champion	8000/tcp	"uvicorn main:app --..."	model-serving-champion
braininvaders_mlops-postgres-1	7 minutes ago	Up 7 minutes (healthy)	postgres:13	5432/tcp	"docker-entrypoint.s..."	postgres
braininvaders_mlops-redis-1	7 minutes ago	Up 7 minutes (healthy)	redis:7.2-bookworm	6379/tcp	"docker-entrypoint.s..."	redis
minio-server	8 minutes ago	Up 5 minutes (healthy)	minio/minio:latest	0.0.0.0:9000-9001->9000-9001/tcp	"/usr/bin/docker-ent..."	minio-server
mlflow-server	8 minutes ago	Up 7 minutes (unhealthy)	braininvaders_mlops-mlflow-server	0.0.0.0:5001->5000/tcp	"mlflow server --hos..."	mlflow-server
nginx-load-balancer	8 minutes ago	Up 7 minutes (healthy)	nginx:latest	0.0.0.0:80->80/tcp	"/docker-entrypoint..."	nginx-load-balancer
postgres-mlflow	8 minutes ago	Up 8 minutes (healthy)	postgres:latest	5432/tcp	"docker-entrypoint.s..."	postgres-mlflow

We can see that mlflow-server is unhealthy, which is unbeknownst to me why. I did every single thing to fix this but literally nothing worked. It doesn't seem to affect it negatively constantly but there were some weird 502 Errors in Nginx service, maybe it was because of this.

Data preparation

airflow-webserver container's terminal:

- `airflow dags trigger p300_full_ingest_and_process_pipeline (2 repeats)`
- `airflow dags unpause p300_full_ingest_and_process_pipeline`
- `airflow dags list-runs -d p300_full_ingest_and_process_pipeline`

namely,

- (airflow)airflow dags list-runs -d p300_full_ingest_and_process_pipeline

dag_id	run_id	state	execution_date	start_date	end_date
p300_full_ingest_and_proces s_pipeline	manual__2025-06-03T13:49:4 6+00:00	running	2025-06-03T13:49:46+00:00	2025-06-03T13:49:47.648487+	
p300_full_ingest_and_proces s_pipeline	manual__2025-06-03T13:48:5 1+00:00	success	2025-06-03T13:48:51+00:00	2025-06-03T13:48:56.816589+	2025-06-03T13:49:22.511258+

Training and deploying challenger model (v1)


- docker compose run --rm --build training-job python train_model.py --
training_subjects_percentage 0.5

Logs of successful alias setting and logging the model version to MLFlow (fitted only 1-5 subjects out of 10 available in MinIO):

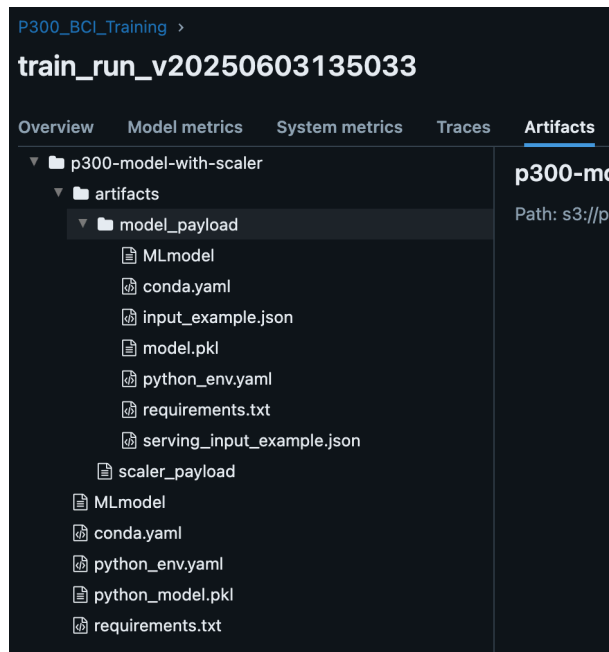
```
mlflow tracking uri confirmed by client: http://mlflow-server:5000
MLflow experiment 'P300_BCI_Training' created with id: 1
starting training pipeline for 5 subjects: [1, 2, 3, 4, 5]
  downloading processed data for subject 01 from p300-processed-features
  downloading processed data for subject 02 from p300-processed-features
  downloading processed data for subject 03 from p300-processed-features
  downloading processed data for subject 04 from p300-processed-features
  downloading processed data for subject 05 from p300-processed-features
combined data: x=(3081, 1120), labels: [1 2]
internal train set: (2464, 1120), internal validation set: (617, 1120)
mlflow run started. run id: c50239584e86450ca190d18640058c6d
  training LDA model
```

```
mlflow: model signature defined explicitly.
Downloading artifacts: 100%| 7/7 [00:00<00:00, 5336.26it/s]
Downloading artifacts: 100%| 1/1 [00:00<00:00, 2159.79it/s]
2025/06/03 13:50:40 WARNING mlflow.models.model: Model logged without a signature and input example. Please set 'input_example' paramet
er when logging the model to auto infer the model signature.
Successfully registered model 'P300-Classifler'.
2025/06/03 13:50:41 INFO mlflow.store.model_registry.abstract_store: Waiting up to 300 seconds for model version to finish creation. Mo
del name: P300-Classifler, version 1
Created version '1' of model 'P300-Classifler'.
mlflow: model and scaler logged. registered 'P300-Classifler' version '1'.
setting alias 'challenger' for model 'P300-Classifler' version '1'.
alias 'challenger' set successfully.
🔗View run train_run_v20250603135033 at: http://mlflow-server:5000/#/experiments/1/runs/c50239584e86450ca190d18640058c6d
🟢View experiment at: http://mlflow-server:5000/#/experiments/1
finished training pipeline. run id: c50239584e86450ca190d18640058c6d
```

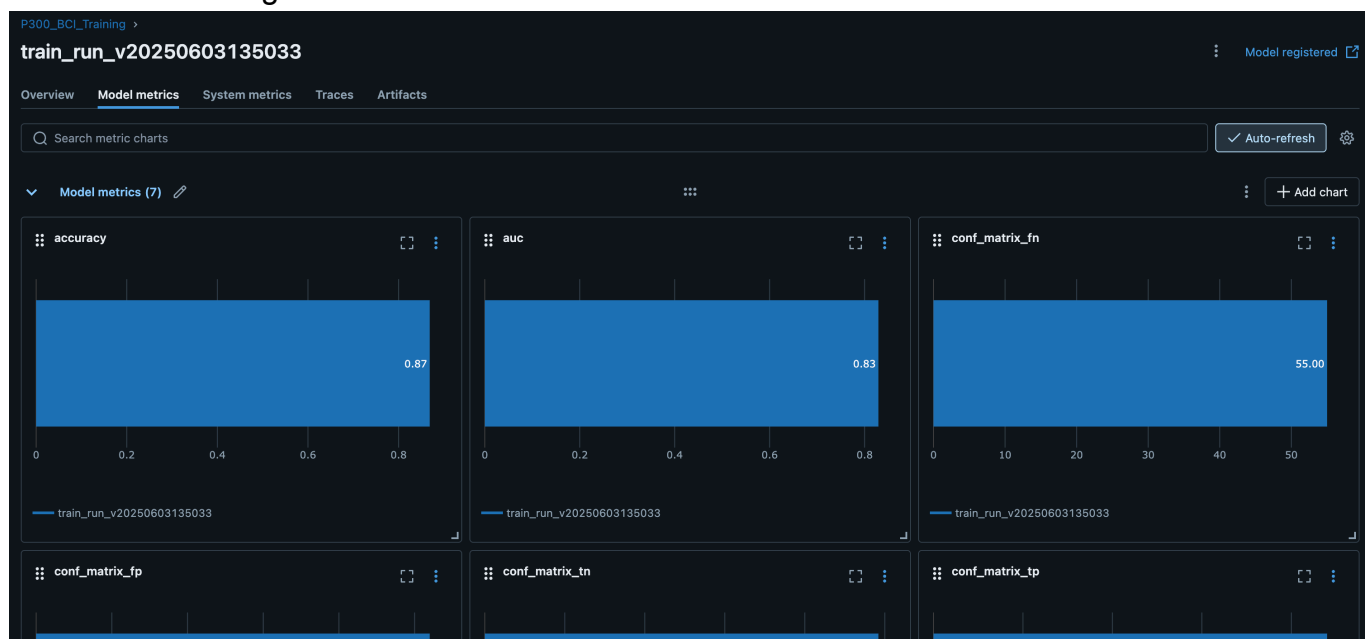
MLflow UI showing Version 1 with the "challenger" alias:

Name 	Latest version	Aliased versions
P300-Classifler	Version 1	@ challenger : Version 1

The artifact structure for this run, containing the bundled model and scaler:



Associated training metrics:



Verifying challenger-v1

- `docker compose restart model-serving-challenger`

We restarted the `model-serving-challenger` service to load the new model version associated with the "challenger" alias.

Logs confirmed successful loading of Version 1 by the challenger service

```
braininvaders_mlops-model-serving-challenger-1
4b2a5e698569 braininvaders_mlops-model-serving-challenger:latest
STATUS
Running (17 seconds ago)

Logs Inspect Bind mounts Exec Files Stats
INFO: Started server process [1]
/usr/local/lib/python3.9/site-packages/mlflow/pyfunc/utils/data_validation.py:186: UserWarning: Add type hints to the 'predict' method to enable data validation and automatic signature inference during model logging. Check https://mlflow.org/docs/latest/model/python_model.html#python-model for more details.
color_warning(
INFO: Started server process [1]
INFO: Waiting for application startup.
INFO: Application startup complete.
INFO: Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
fastapi application startup. loading model: models:/P300-Classifier@challenger
attempting to load model from MLflow registry using URI: models:/P300-Classifier@challenger
resolved alias 'challenger' to version '1' of model 'P300-Classifier'.
downloading entire pyfunc model (scaler + classifier) artifact from: s3://p300-mlflow-artifacts/1/c50239584e86450ca190d18640058c6
scaler
contents of 'artifacts' subdir (/tmp/tmp3b0p55hw/p300-model-with-scaler/artifacts): ['scaler_payload', 'model_payload']
scikit-learn model loaded successfully from /tmp/tmp3b0p55hw/p300-model-with-scaler/artifacts/model_payload
scaler loaded successfully from /tmp/tmp3b0p55hw/p300-model-with-scaler/artifacts/scaler_payload
successfully loaded model version 1 (from models:/P300-Classifier@challenger) and scaler.
INFO: 127.0.0.1:59742 - "GET /health HTTP/1.1" 200 OK
INFO: 127.0.0.1:52778 - "GET /health HTTP/1.1" 200 OK
```

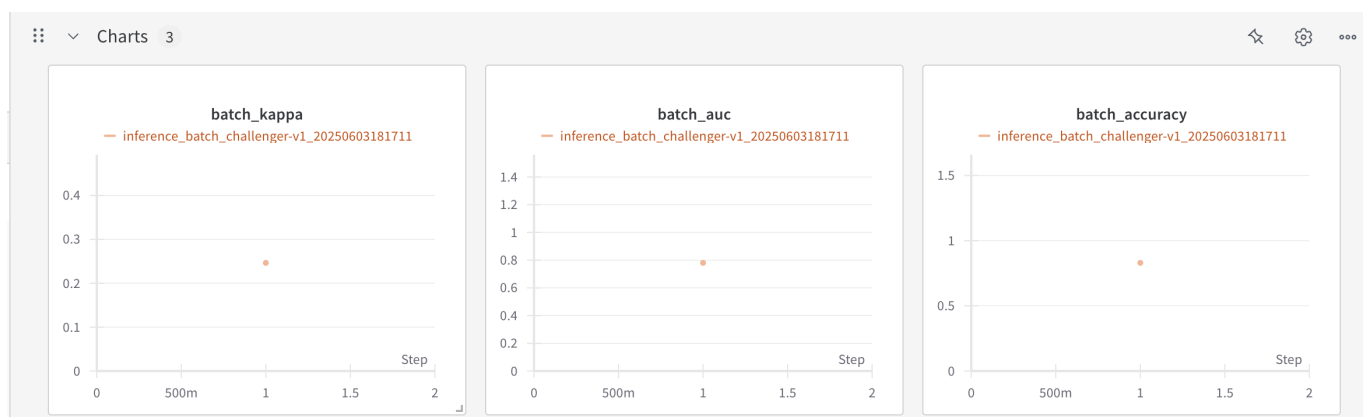
We then ran the application simulator, directing it to the challenger's endpoint via Nginx:

- python application-simulator/application_simulator.py 6 7 --model_alias challenger

The simulator logs and W&B metrics confirmed evaluation against the challenger-v1:

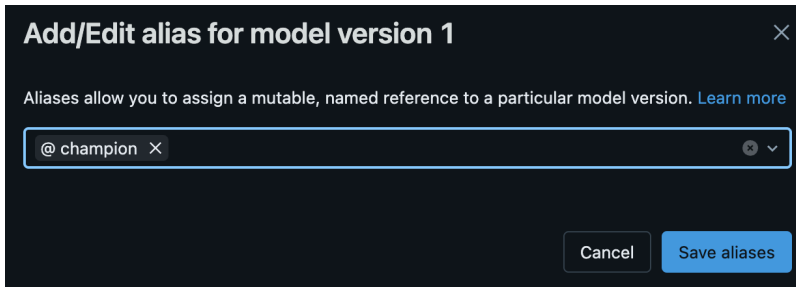
```
targeting challenger model (version: 1) at: http://localhost/predict-challenger/
wandb: Currently logged in as: xphoenixua to https://api.wandb.ai. Use `wandb login --relogin` to force relogin
wandb: WARNING Using a boolean value for 'reinit' is deprecated. Use 'return_previous' or 'finish_previous' instead.
wandb: Tracking run with wandb version 0.19.11
wandb: Run data is saved locally in /Users/user/braininvaders_mlops/wandb/run-20250603_170250-1fkpemxq
wandb: Run `wandb offline` to turn off syncing.
wandb: Syncing run inference_batch_challenger-v1_20250603170250
wandb: ★ View project at https://wandb.ai/xphoenixua/p300-bci-mlops
wandb: 🚀 View run at https://wandb.ai/xphoenixua/p300-bci-mlops/runs/1fkpemxq
simulating predictions for batch of subjects: [6, 7]
```

W&B charts for this evaluation:



Promoting challenger-v1 to champion-v1

In the MLflow UI, we manually updated the aliases for "P300-Classifer" Version 1: removed the "challenger" alias and added the "champion" alias.



We then restarted the champion service replicas. These instances now load Model Version 1 as it holds the "champion" alias.

- `docker compose restart model-serving-champion`

Verifying champion-v1 model

We ran the application simulator, targeting the champion endpoint:

- `python application-simulator/application_simulator.py 9 10 --model_alias champion`

Simulator logs confirmed evaluation against the new champion-v1. Each prediction response (POST verbs in the logs) was redistributed among the `model-serving-champion` replicas, demonstrating that Nginx was distributing load across them:

- `docker compose run --rm --build training-job python train_model.py --training_subjects_percentage 0.75`

Logs confirmed registration of Version 2 and assignment of the "challenger" alias. Version 1 remained as "champion".

```
Registered model 'P300-Classifer' already exists. Creating a new version of this model...
2025/06/03 15:23:37 INFO mlflow.store.model_registry.abstract_store: Waiting up to 300 seconds for model version to finish creation. Model name: P300-Classifer, version 2
Created version '2' of model 'P300-Classifer'.
mlflow: model and scaler logged. registered 'P300-Classifer' version '2'.
setting alias 'challenger' for model 'P300-Classifer' version '2'.
alias 'challenger' set successfully.
🔗 View run train_run_v20250603152318 at: http://mlflow-server:5000/#/experiments/1/runs/c7696f7a9f9847eb92bb3d5d49a87d1c
🟢 View experiment at: http://mlflow-server:5000/#/experiments/1
finished training pipeline. run id: c7696f7a9f9847eb92bb3d5d49a87d1c
```

We restarted the `model-serving-challenger` service and it now loaded model-v2.

- `docker compose restart model-serving-challenger`

```
INFO: Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
fastapi application startup. loading model: models:P300-Classifer@challenger
attempting to load model from MLflow registry using URI: models:P300-Classifer@challenger
resolved alias 'challenger' to version '2' of model 'P300-Classifer'.
downloading entire pyfunc model (scaler + classifier) artifact from: s3://p300-mlflow-artifacts/1/c7696f7a9f9847eb92bb3d5d49a87d1c
scaler
contents of 'artifacts' subdir (/tmp/tmpcdjn65g6/p300-model-with-scaler/artifacts): ['scaler_payload', 'model_payload']
scikit-learn model loaded successfully from /tmp/tmpcdjn65g6/p300-model-with-scaler/artifacts/model_payload
scaler loaded successfully from /tmp/tmpcdjn65g6/p300-model-with-scaler/artifacts/scaler_payload
successfully loaded model version 2 (from models:P300-Classifer@challenger) and scaler.
```

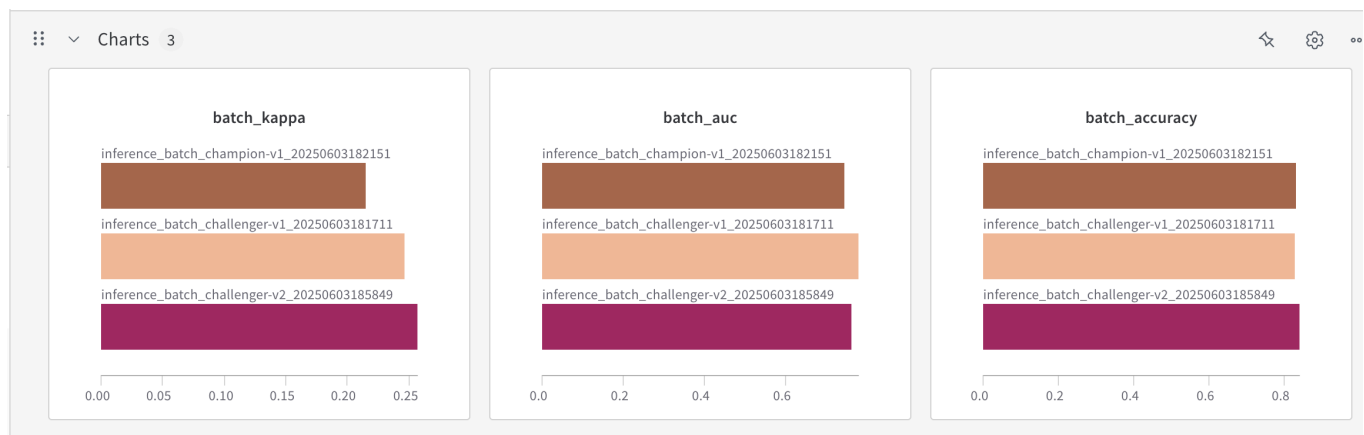
Verifying challenger-v2

We ran the application simulator against the challenger endpoint:

- `python application-simulator/application_simulator.py 9 10 --model_alias challenger`

```
targeting challenger model (version: 2) at: http://localhost/predict-challenger/
wandb: Currently logged in as: xphoenixua to https://api.wandb.ai. Use `wandb login --relogin` to force relogin
wandb: WARNING Using a boolean value for 'reinit' is deprecated. Use 'return_previous' or 'finish_previous' instead.
wandb: Tracking run with wandb version 0.19.11
wandb: Run data is saved locally in /Users/user/braininvaders_mlops/wandb/run-20250603_185850-fqaysj4j
wandb: Run `wandb offline` to turn off syncing.
wandb: Syncing run inference_batch_challenger-v2_20250603185849
wandb: ★ View project at https://wandb.ai/xphoenixua/p300-bci-mlops
wandb: 🚀 View run at https://wandb.ai/xphoenixua/p300-bci-mlops/runs/fqaysj4j
simulating predictions for batch of subjects: [8, 9]
```

The simulator's W&B logs and its output (specifically the `model_version_loaded` field derived from the API response) confirmed that challenger-v2 was being served by the challenger endpoint.



Config

Config parameters are your model's inputs. [Learn more](#)

▼ **Config parameters:** {} 4 keys

serving_api_url_used: <http://localhost/predict-challenger/>

▼ **subject_ids_inferred_in_batch:** [] 2 items

0: 8

1: 9

target_model_type_simulated: "challenger"

target_model_version_simulated: "2"

Finishing the pipeline

- `docker compose -f airflow-compose.yaml down -v`
- `docker compose down -v`