MLOps Zoomcamp 2023

This is the homework for week#2 `02-experiment-tracking` module, for the above course's cohort 2023. We're supposed to practice with the `Green Taxi Trip Records` of the NYC taxi dataset https://www.nyc.gov's "TLC Trip Record Data" for year `2022`.

The problem statement we're solving is to predict the `trip_amount`.

Submit the answers to https://forms.gle/Fy1pvrPEKd4yjz3s6 by

- 1 Jun 2023 (Tuesday), 23:00 CEST (Berlin time)
- 2 Jun 2023 (Wednesday), 05:00 SST (Singapore local time)

Notes for the lessons can be found in

In [ ]:

```
!mlflow --version
```

```
mlflow, version 1.25.0
```

Q1. Install the package

Q1. What's the version of mlflow that you have?

A1. 1.25.0

Q2. Download and preprocess the data

command `python preprocess_data.py --raw_data_path data/raw/ --dest_path data/processed/`

note: I changed line#18 in `train.py` later, so here need to supply output as changed below

- from `./output`
- to `data/processed/`

Q2. So what's the size of the saved DictVectorizer file?

A2. 152 KB (155,648 bytes)

Q3. Train a model with autolog

checklist before launching mlflow ui:

- [ ] remove any .db files
- [ ] remove any numbered files under mlruns/(#), if still exist after previous runs and after killing process
- [ ] split 2 terminal panels

commands

- `mlflow ui --backend-store-uri sqlite:///mlflow.db`
- `ps -A | grep gunicorn` then `kill <process-id>`
- `sudo fuser -k 5000/tcp` to simply kill all processes using port 5000
- `python train.py`

edit train.py to reproduce experiment; git diff on file should show these

```
> #04    import mlflow
> #05    import mlflow.sklearn
> #10    mlflow.set_tracking_uri("sqlite:///mlflow.db")
> #11    mlflow.set_experiment("train-random-forest")
> #19    default="data/processed",
> #23    mlflow.sklearn.autolog()
> #25    with mlflow.start_run():
> #35    mlflow.log_metric("rmse", rmse)
```

Q3: What is the value of the max_depth parameter:

A3: 10

Q4. Tune model hyperparameters

commands

- `python hpo.py`

Q4: What's the best validation RMSE that you got?

A4: 2.45

Q5. Promote the best model to the model registry

commands

- `python register_model.py`

Q5: What is the test RMSE of the best model?

A5: 2.291 (take closest 2.185)