ml-pipeline

CI/CD pipeline for ML model for BigData course

Stack

- Analytics and model training
  - Python 3.x
  - Pandas, NumPy, SkLearn
- Testing
  - unittest + coverage
- Data/Model versioning
  - DVC
- CI/CD
  - Github Actions
- Secrets vault
  - Ansible Vault
- Results storage
  - MySQL
  - Apache Kafka

Dataset

Dataset was collected using information about rides from bike sharing system. The goal is to predict number of rents during specific hour based on time, day, season, wheather and etc.

Link: https://www.kaggle.com/competitions/bike-sharing-demand

Workflow

- Downloaded dataset from Kaggle
- Analyzed and tuned model
- Transformed research notebook into scripts
- Put dataset to S3 using DVC
- Created Dockerfile and docker-compose.yml
- Created piplines usin GitHub Actions
- Added logging of features, predictions and target during functional tests to MySQL database
- Added logging to special Apache Kafka topic

Test results (CD)

```
r_forest_training_1  | INFO:root:Training...
r_forest_training_1  | INFO:root:Train MSE 234.4049369200735 | Val MSE 1559.864693801653
r_forest_training_1  | ....
r_forest_training_1  | ----------------------------------------------------------------------
```

```
r_forest_training_1  | Ran 4 tests in 1.388s
r_forest_training_1  |
r_forest_training_1  | OK
r_forest_training_1  | INFO:root:Training...
r_forest_training_1  | .
r_forest_training_1  | ----------------------------------------------------------------------
r_forest_training_1  | Ran 1 test in 4.223s
r_forest_training_1  |
r_forest_training_1  | OK
r_forest_training_1  | Name                                Stmts   Miss  Cover   Missing
r_forest_training_1  | -----------------------------------------------------------------
r_forest_training_1  | src/preprocess.py                      21      0   100%
r_forest_training_1  | src/trainer.py                         68     20    71%   46-47, 85-8
r_forest_training_1  | src/unit_tests/test_preprocess.py      30      0   100%
r_forest_training_1  | src/unit_tests/test_training.py        19      0   100%
r_forest_training_1  | -----------------------------------------------------------------
r_forest_training_1  | TOTAL                                 138     20    86%
ml-pipeline_r_forest_training_1 exited with code 0
```