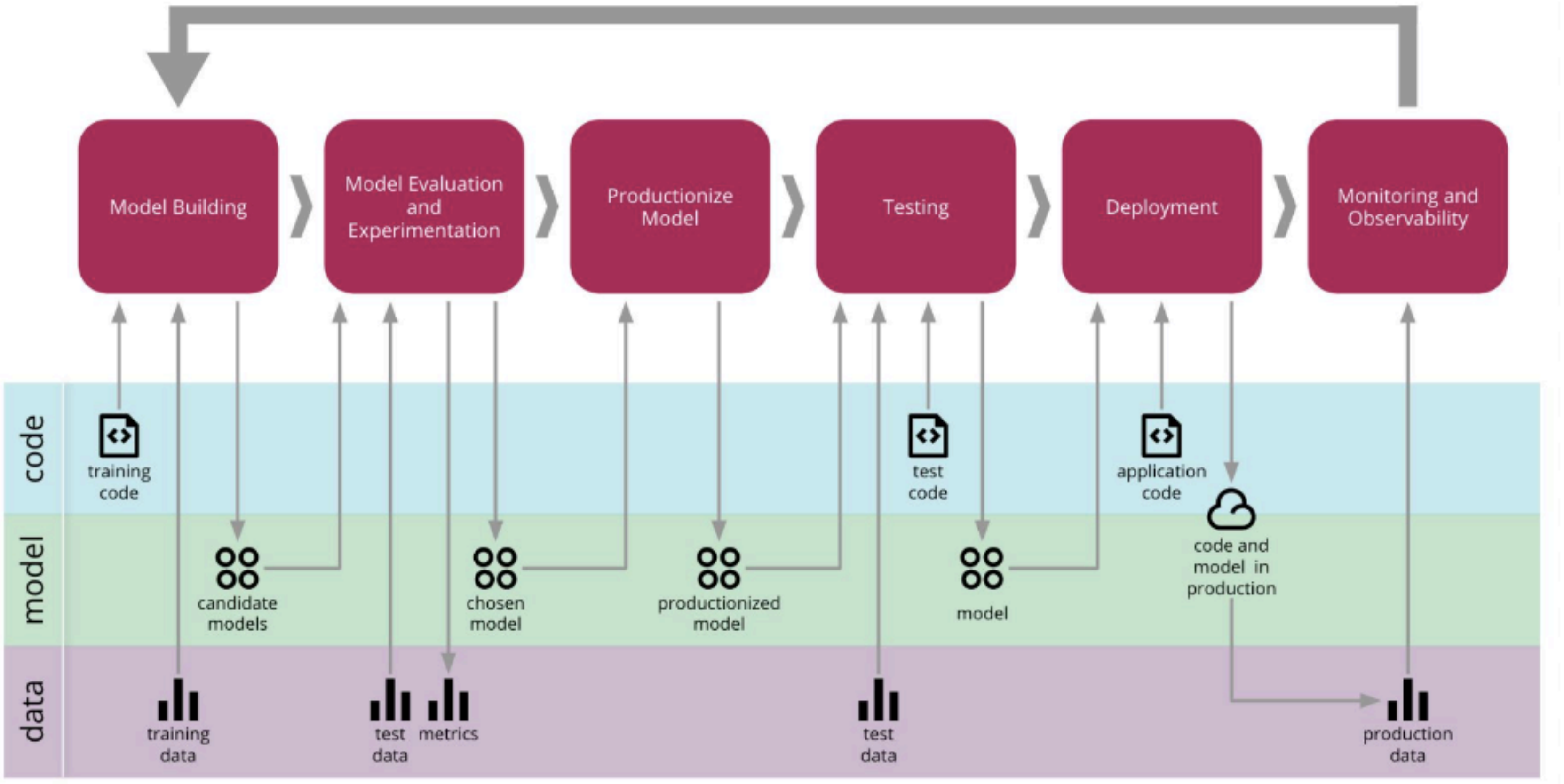
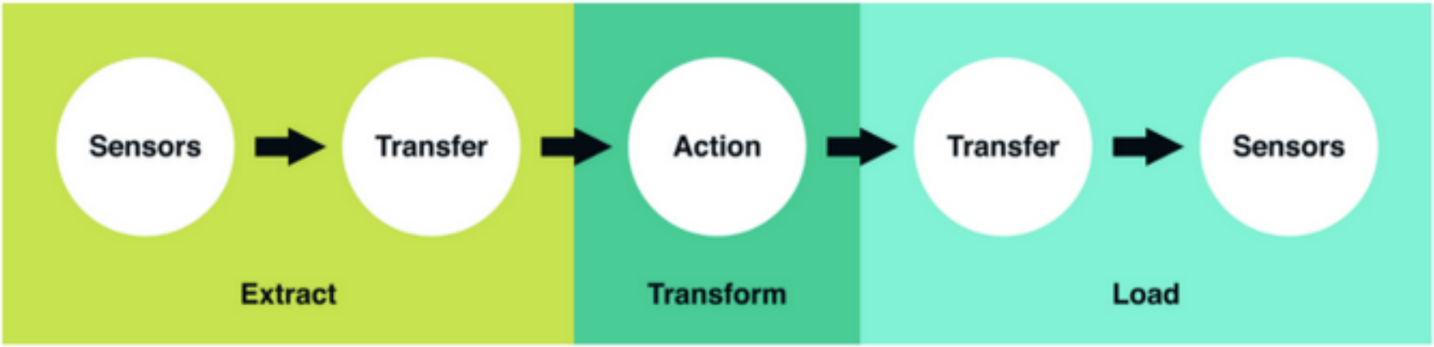
Workflows e Rastreamento

Luciano Barbosa

Workflow de Machine Learning

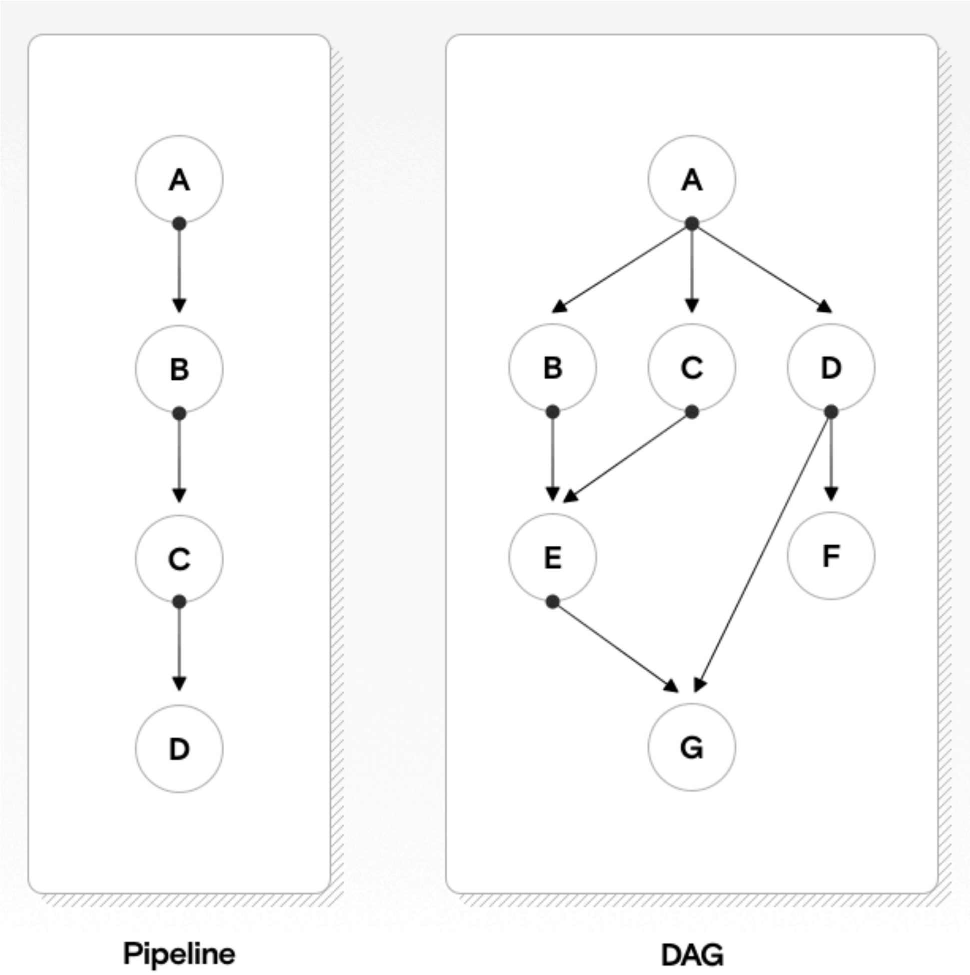


ETL



Source: Vineet Goel's "Why Robinhood uses Airflow?" Medium Post

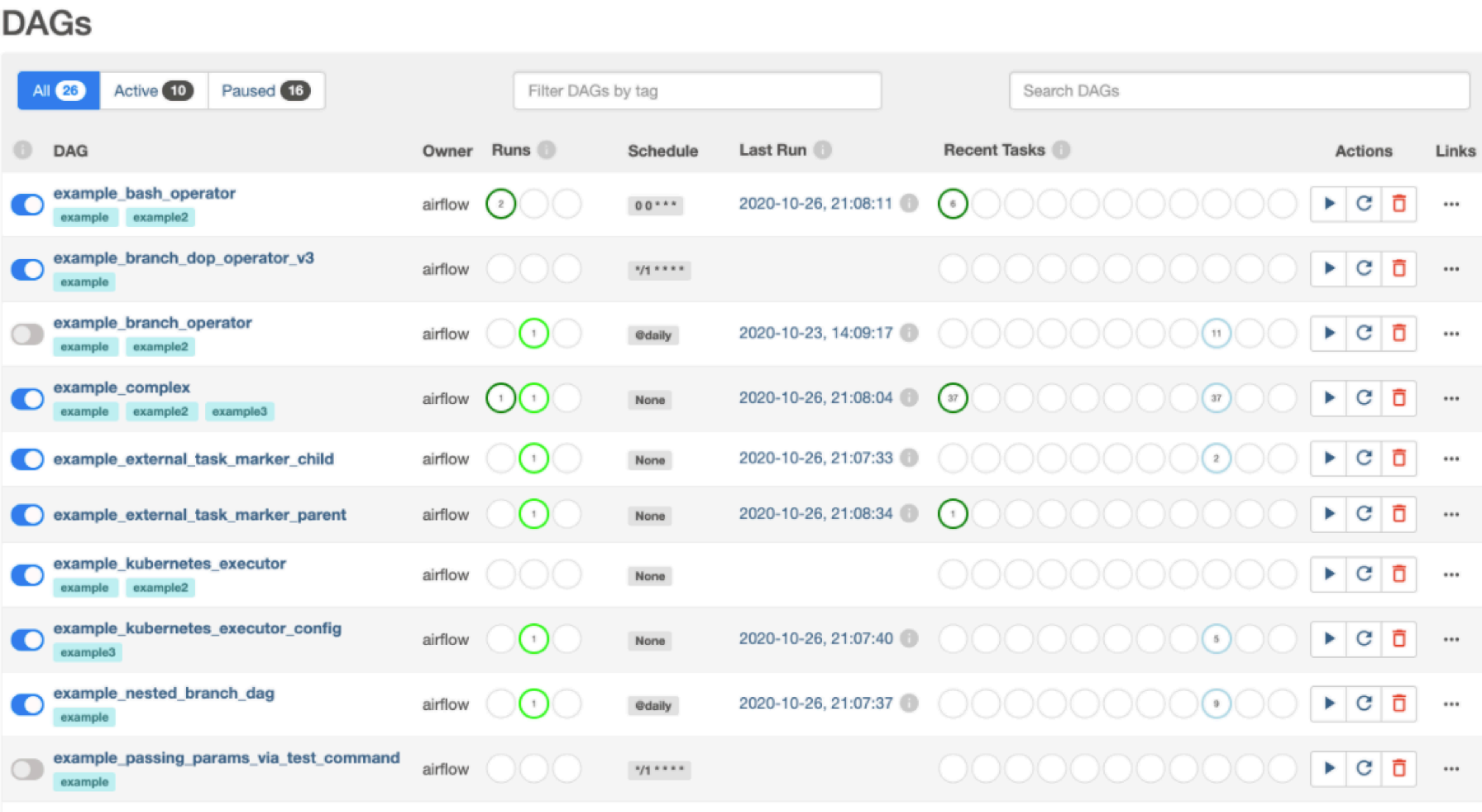
Directed Acyclic Graph - DAG



Airflow: Operadores

- Sensors: inicia a tarefa
  - Agendado no tempo
  - Dados de entrada disponíveis
- Operators: executa a tarefa
  - Python, docker, mysql, s3, email, http, spark
- Transfers: transfere dados de um lugar para outro
  - google_api_to_s3_transfer

Airflow



DAGs

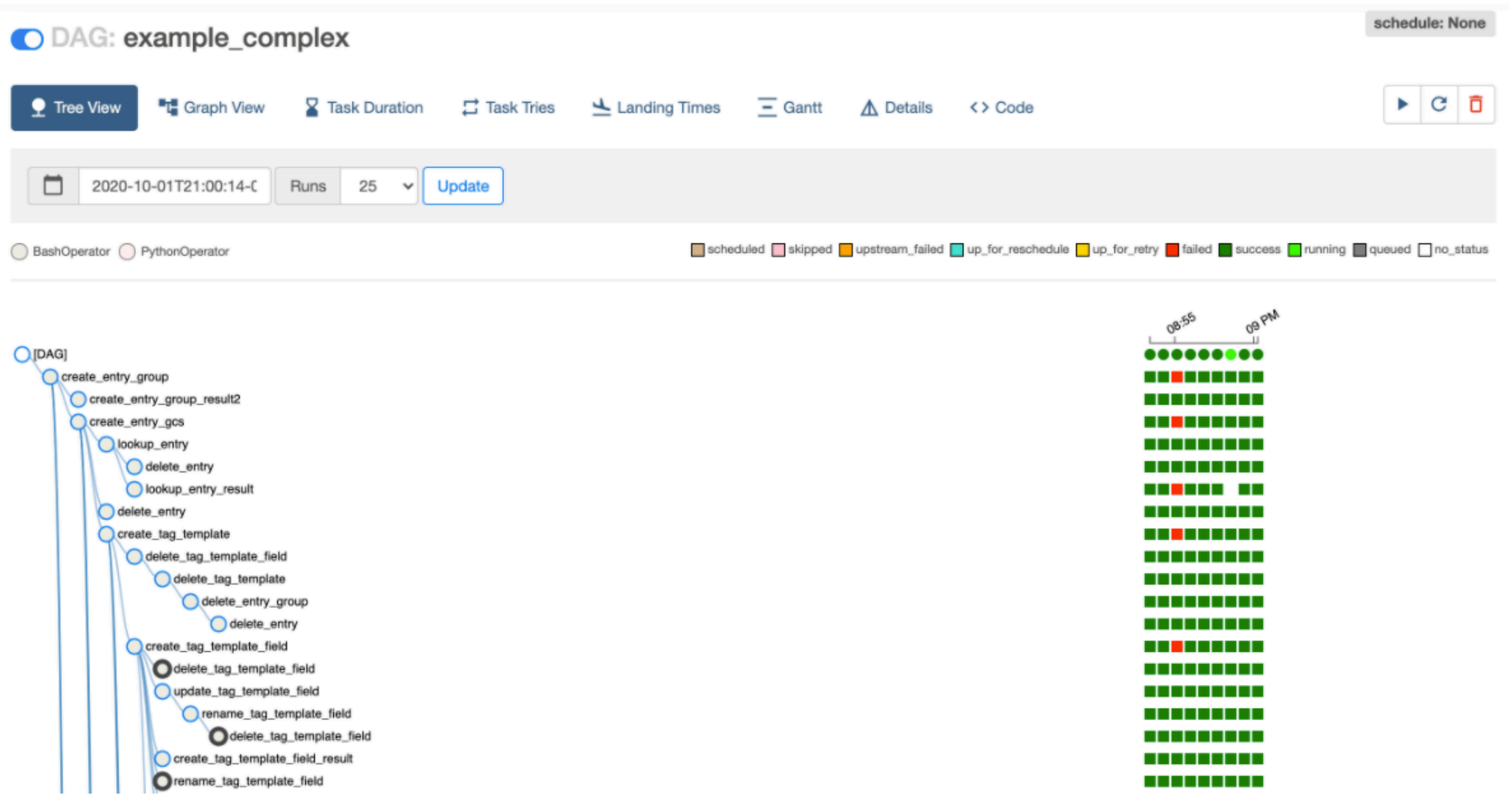
All 26 | Active 10 | Paused 16

Filter DAGs by tag

Search DAGs

	DAG	Owner	Runs	Schedule	Last Run	Recent Tasks	Actions	Links
●	example_bash_operator example example2	airflow	2	0 0 * * *	2020-10-26, 21:08:11	6	▶ ↻ 🗑	•••
●	example_branch_dop_operator_v3 example	airflow		*/1 * * * *			▶ ↻ 🗑	•••
○	example_branch_operator example example2	airflow	1	@daily	2020-10-23, 14:09:17	11	▶ ↻ 🗑	•••
●	example_complex example example2 example3	airflow	1 1	None	2020-10-26, 21:08:04	37 37	▶ ↻ 🗑	•••
●	example_external_task_marker_child	airflow	1	None	2020-10-26, 21:07:33	2	▶ ↻ 🗑	•••
●	example_external_task_marker_parent	airflow	1	None	2020-10-26, 21:08:34	1	▶ ↻ 🗑	•••
●	example_kubernetes_executor example example2	airflow		None			▶ ↻ 🗑	•••
●	example_kubernetes_executor_config example3	airflow	1	None	2020-10-26, 21:07:40	5	▶ ↻ 🗑	•••
●	example_nested_branch_dag example	airflow	1	@daily	2020-10-26, 21:07:37	9	▶ ↻ 🗑	•••
○	example_passing_params_via_test_command example	airflow		*/1 * * * *			▶ ↻ 🗑	•••

Airflow



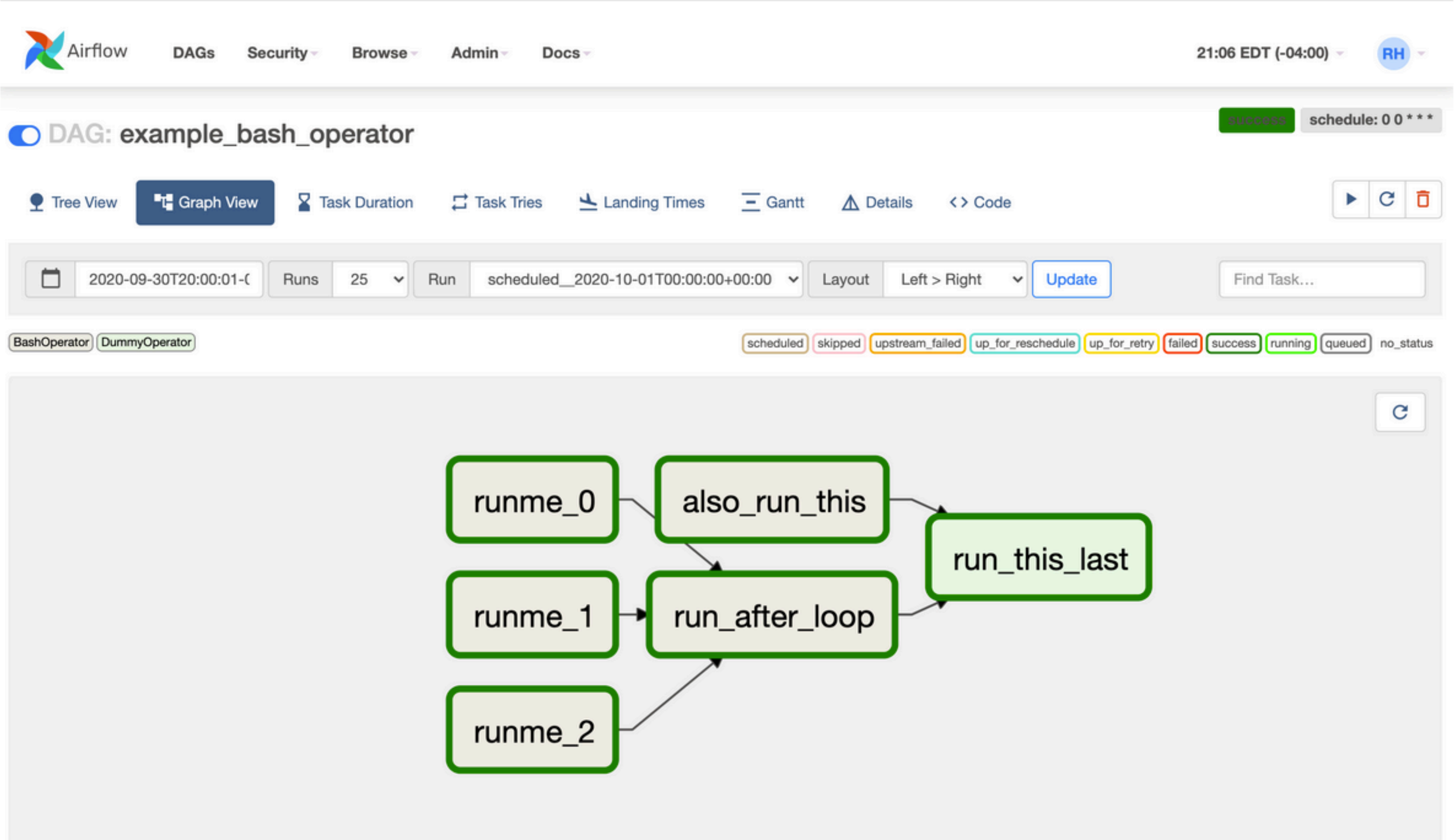
DAG: example_complex

schedule: None

2020-10-01T21:00:14-(

Runs 25

Airflow



DAG: example_bash_operator

success

schedule: 0 0 * * *

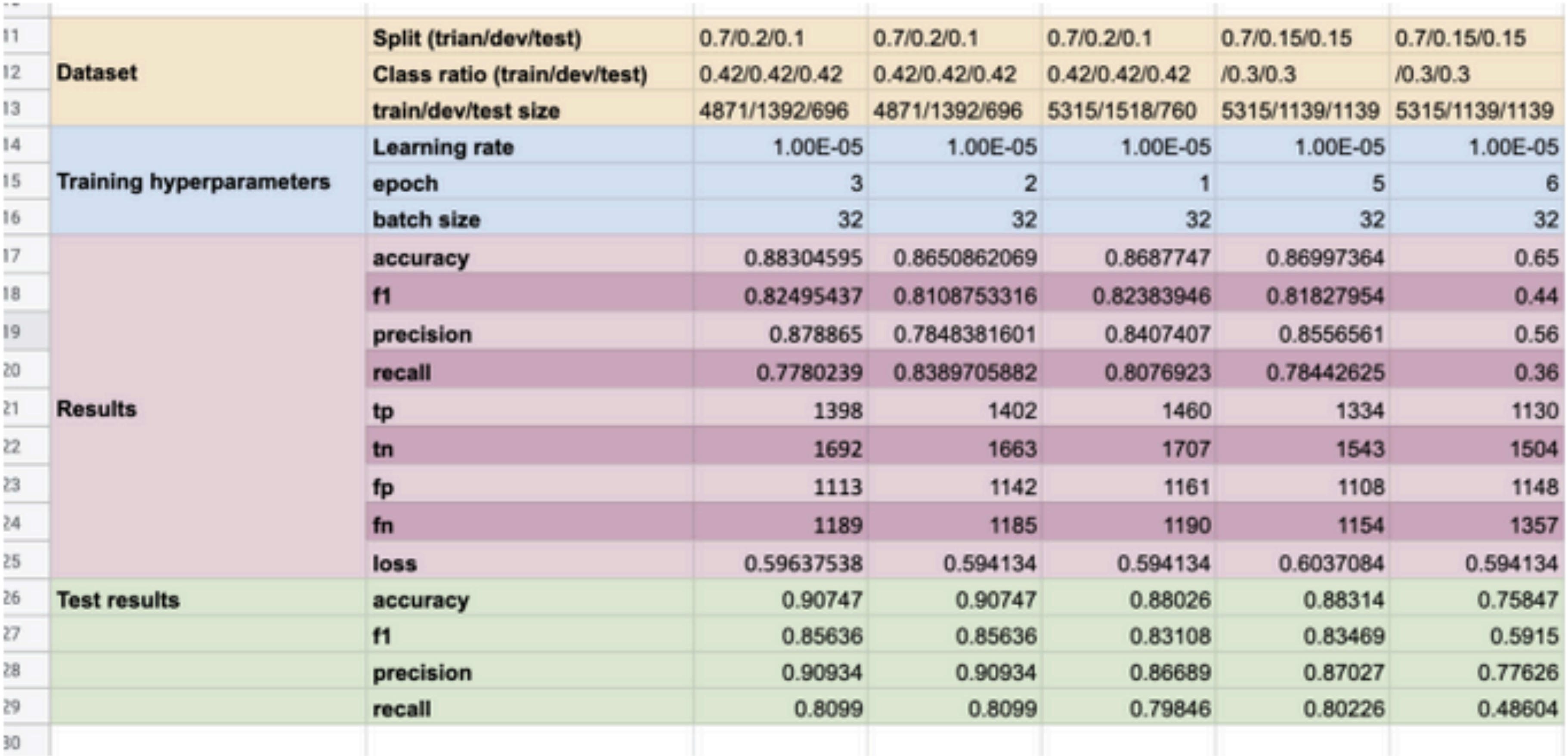
2020-09-30T20:00:01-(

Runs 25

Run scheduled__2020-10-01T00:00:00+00:00

Layout Left > Right

Rastreando Experimentos com Planilhas

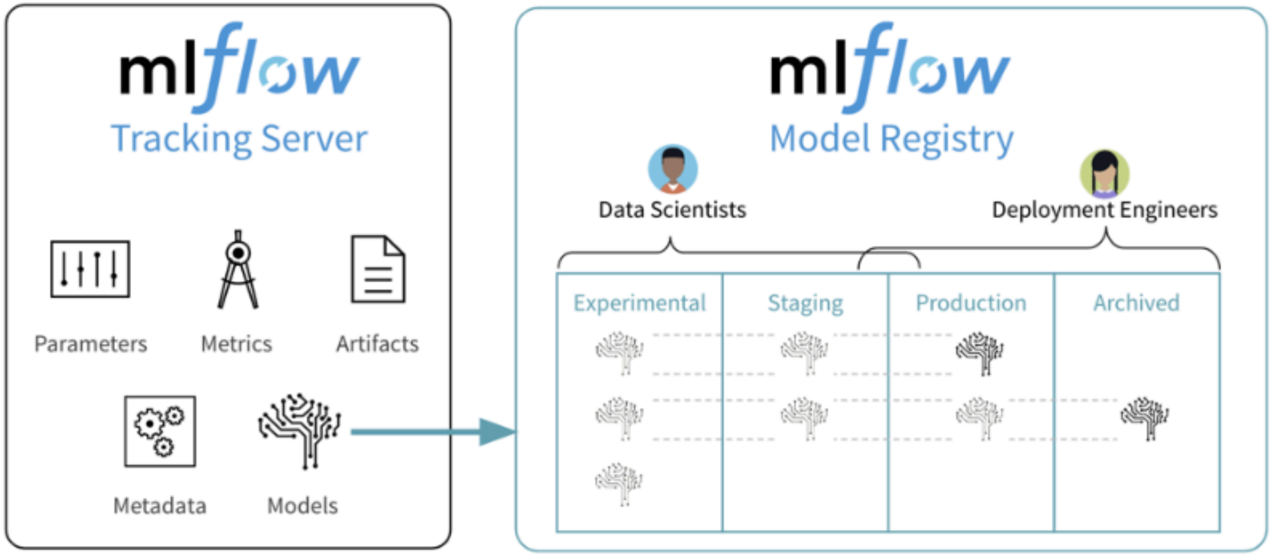


11	Dataset	Split (trian/dev/test)	0.7/0.2/0.1	0.7/0.2/0.1	0.7/0.2/0.1	0.7/0.15/0.15	0.7/0.15/0.15
12		Class ratio (train/dev/test)	0.42/0.42/0.42	0.42/0.42/0.42	0.42/0.42/0.42	/0.3/0.3	/0.3/0.3
13		train/dev/test size	4871/1392/696	4871/1392/696	5315/1518/760	5315/1139/1139	5315/1139/1139
14	Training hyperparameters	Learning rate	1.00E-05	1.00E-05	1.00E-05	1.00E-05	1.00E-05
15		epoch	3	2	1	5	6
16		batch size	32	32	32	32	32
17	Results	accuracy	0.88304595	0.8650862069	0.8687747	0.86997364	0.65
18		f1	0.82495437	0.8108753316	0.82383946	0.81827954	0.44
19		precision	0.878865	0.7848381601	0.8407407	0.8556561	0.56
20		recall	0.7780239	0.8389705882	0.8076923	0.78442625	0.36
21		tp	1398	1402	1460	1334	1130
22		tn	1692	1663	1707	1543	1504
23		fp	1113	1142	1161	1108	1148
24		fn	1189	1185	1190	1154	1357
25		loss	0.59637538	0.594134	0.594134	0.6037084	0.594134
26	Test results	accuracy	0.90747	0.90747	0.88026	0.88314	0.75847
27		f1	0.85636	0.85636	0.83108	0.83469	0.5915
28		precision	0.90934	0.90934	0.86689	0.87027	0.77626
29		recall	0.8099	0.8099	0.79846	0.80226	0.48604
30							

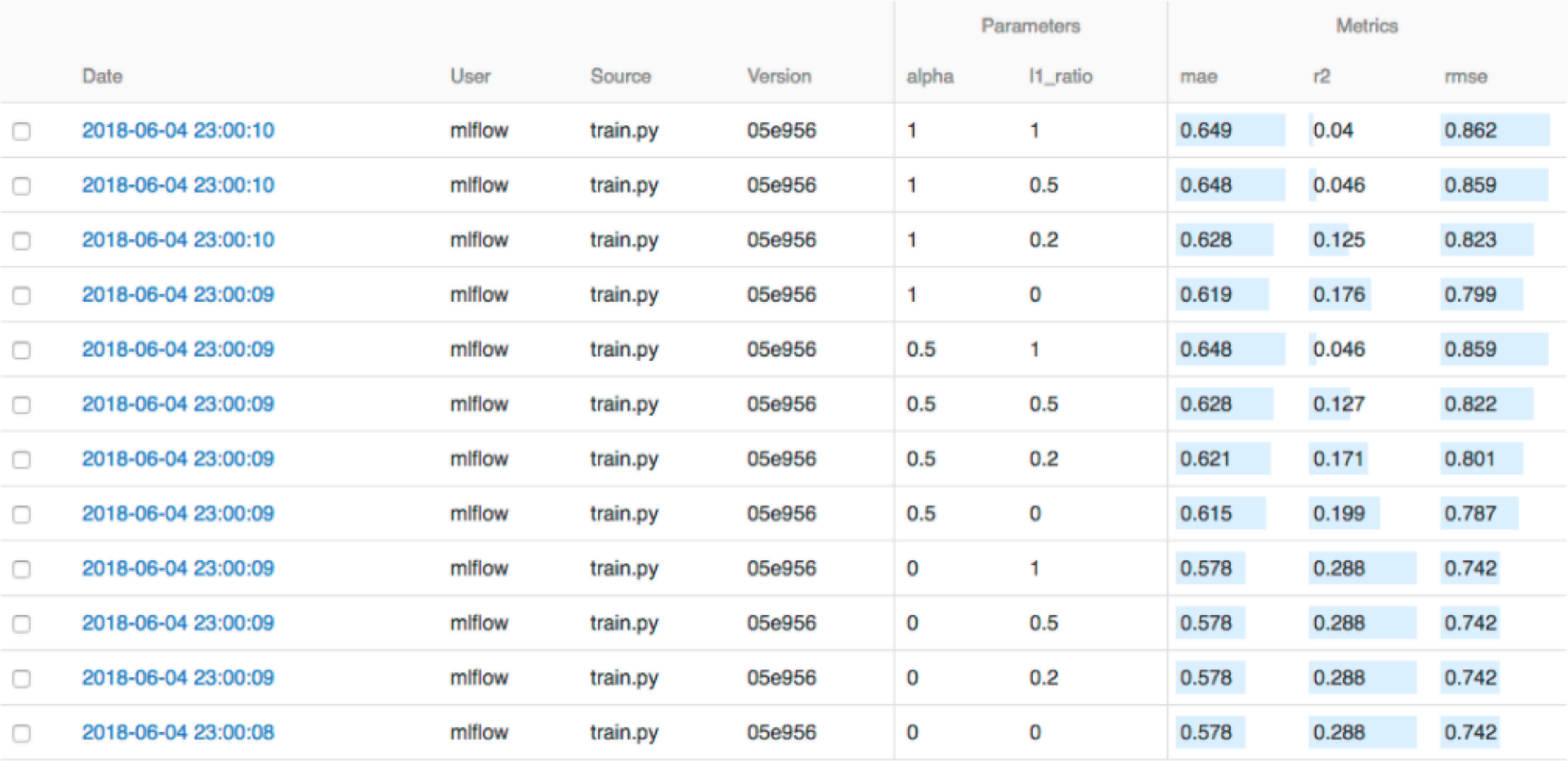
Rastreamento de Resultados

MLflow

- Ferramenta para rastreamento automático de resultados

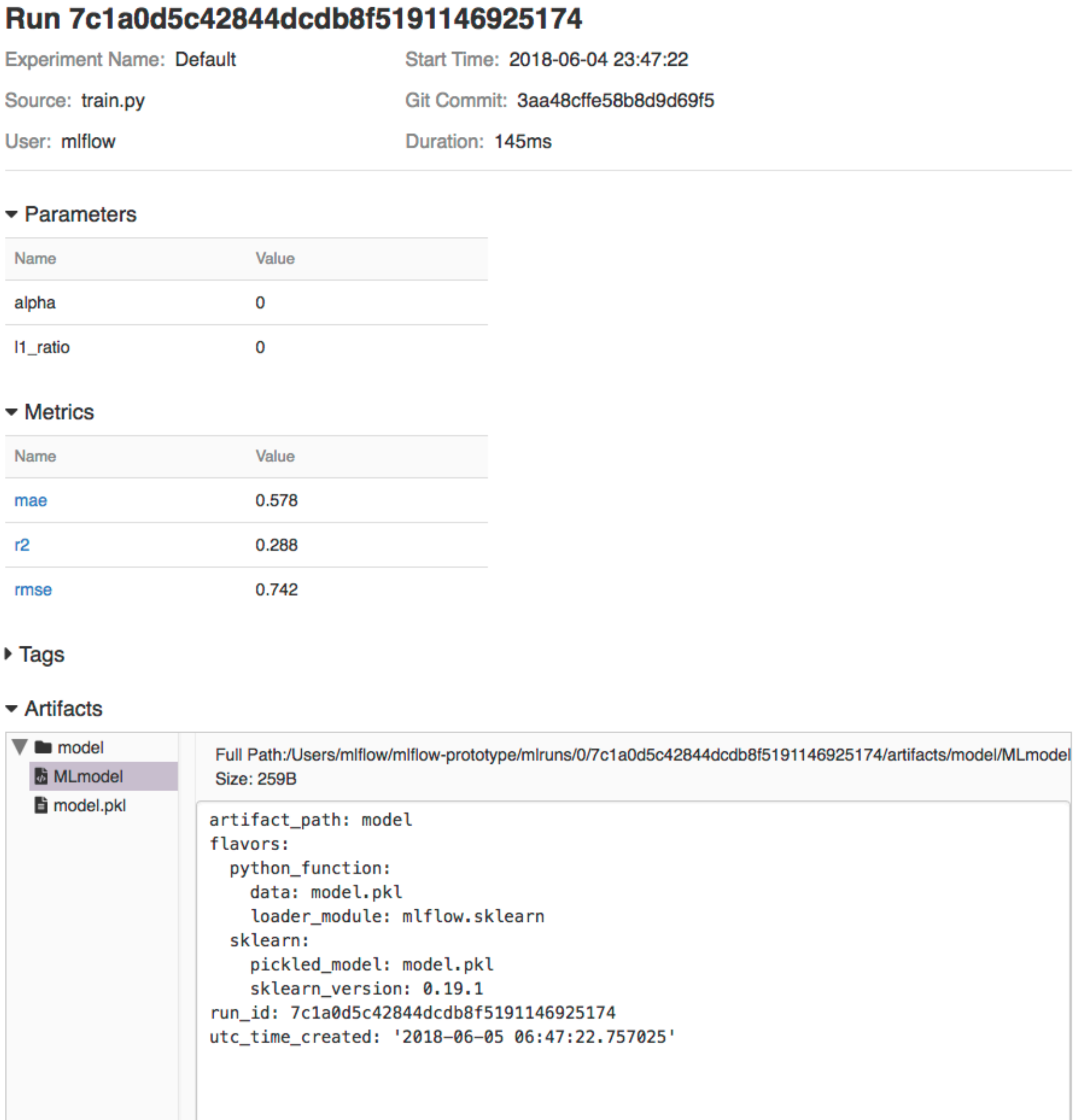


Comparando Modelos



					Parameters		Metrics		
	Date	User	Source	Version	alpha	l1_ratio	mae	r2	rmse
□	2018-06-04 23:00:10	mlflow	train.py	05e956	1	1	0.649	0.04	0.862
□	2018-06-04 23:00:10	mlflow	train.py	05e956	1	0.5	0.648	0.046	0.859
□	2018-06-04 23:00:10	mlflow	train.py	05e956	1	0.2	0.628	0.125	0.823
□	2018-06-04 23:00:09	mlflow	train.py	05e956	1	0	0.619	0.176	0.799
□	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	1	0.648	0.046	0.859
□	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	0.5	0.628	0.127	0.822
□	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	0.2	0.621	0.171	0.801
□	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	0	0.615	0.199	0.787
□	2018-06-04 23:00:09	mlflow	train.py	05e956	0	1	0.578	0.288	0.742
□	2018-06-04 23:00:09	mlflow	train.py	05e956	0	0.5	0.578	0.288	0.742
□	2018-06-04 23:00:09	mlflow	train.py	05e956	0	0.2	0.578	0.288	0.742
□	2018-06-04 23:00:08	mlflow	train.py	05e956	0	0	0.578	0.288	0.742

Tela de Rastreamento



Run 7c1a0d5c42844dcdb8f5191146925174

Experiment Name: Default

Start Time: 2018-06-04 23:47:22

Source: train.py

Git Commit: 3aa48cffe58b8d9d69f5

User: mlflow

Duration: 145ms

▾ Parameters

Name	Value
alpha	0
l1_ratio	0

▾ Metrics

Name	Value
mae	0.578
r2	0.288
rmse	0.742

▸ Tags

▾ Artifacts

- ▼ model
  - MLmodel
  - model.pkl

Full Path:/Users/mlflow/mlflow-prototype/mlruns/0/7c1a0d5c42844dcdb8f5191146925174/artifacts/model/MLmodel

Size: 259B

```
artifact_path: model
flavors:
  python_function:
    data: model.pkl
    loader_module: mlflow.sklearn
  sklearn:
    pickled_model: model.pkl
    sklearn_version: 0.19.1
run_id: 7c1a0d5c42844dcdb8f5191146925174
utc_time_created: '2018-06-05 06:47:22.757025'
```