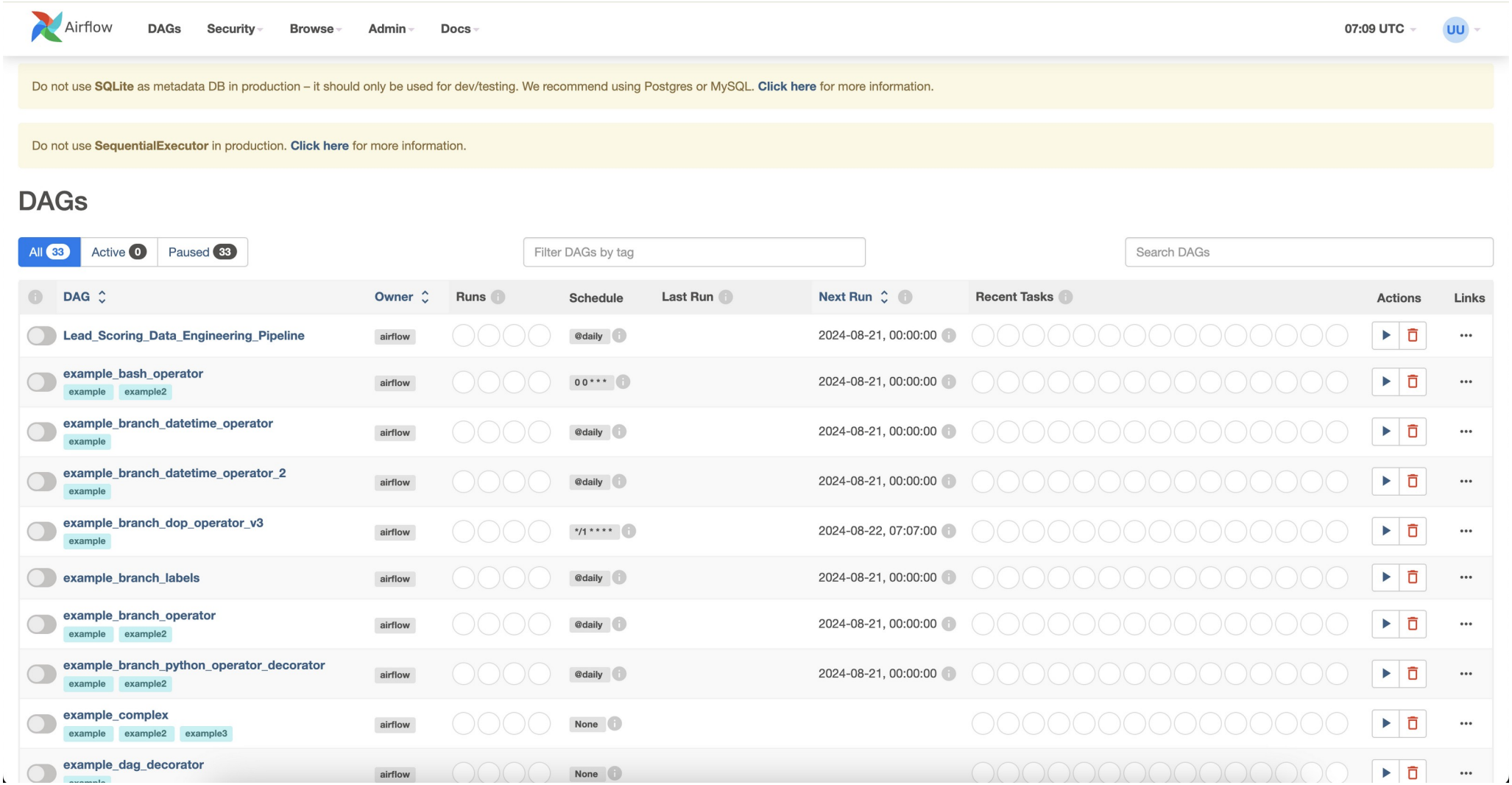
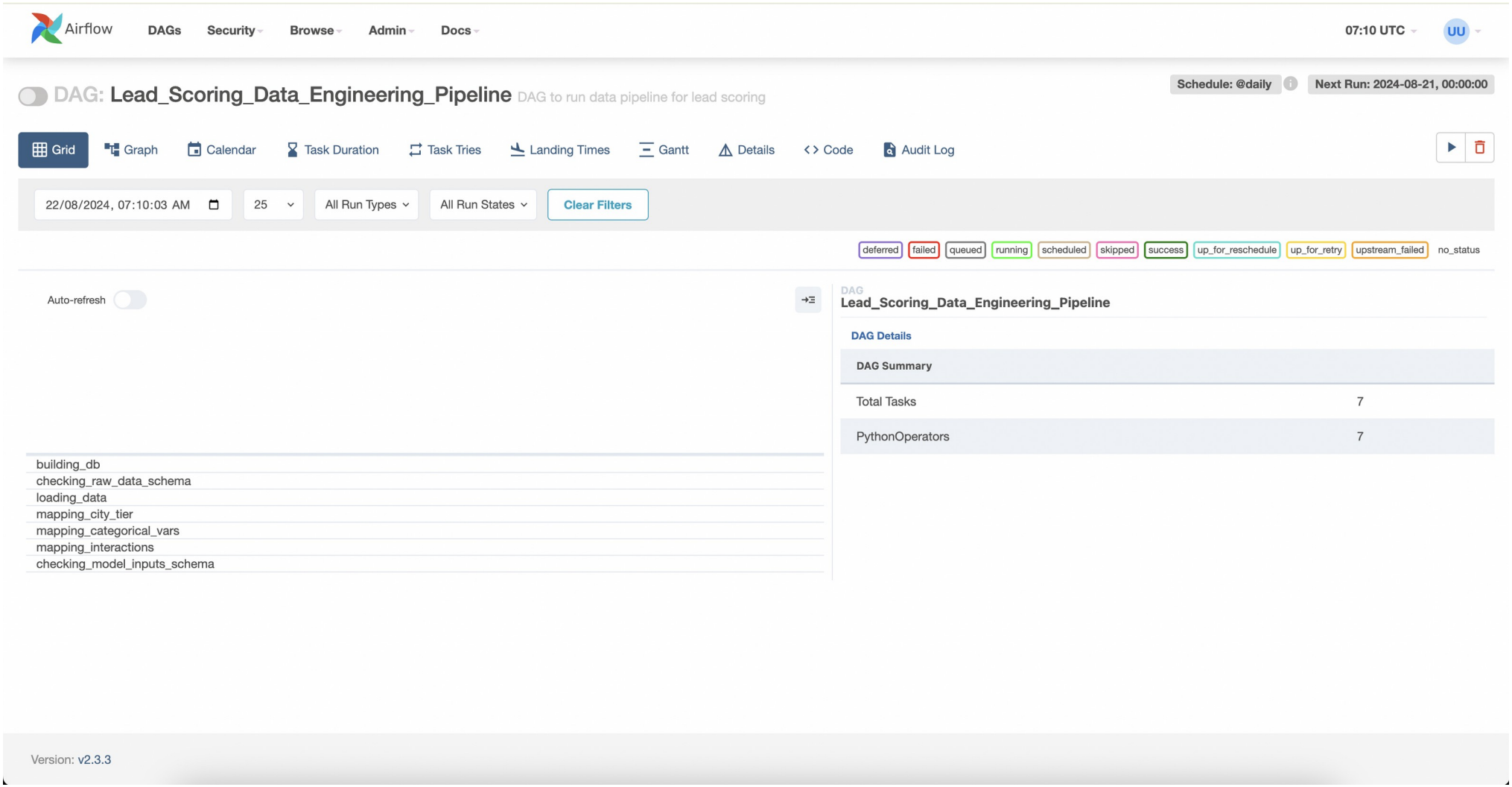
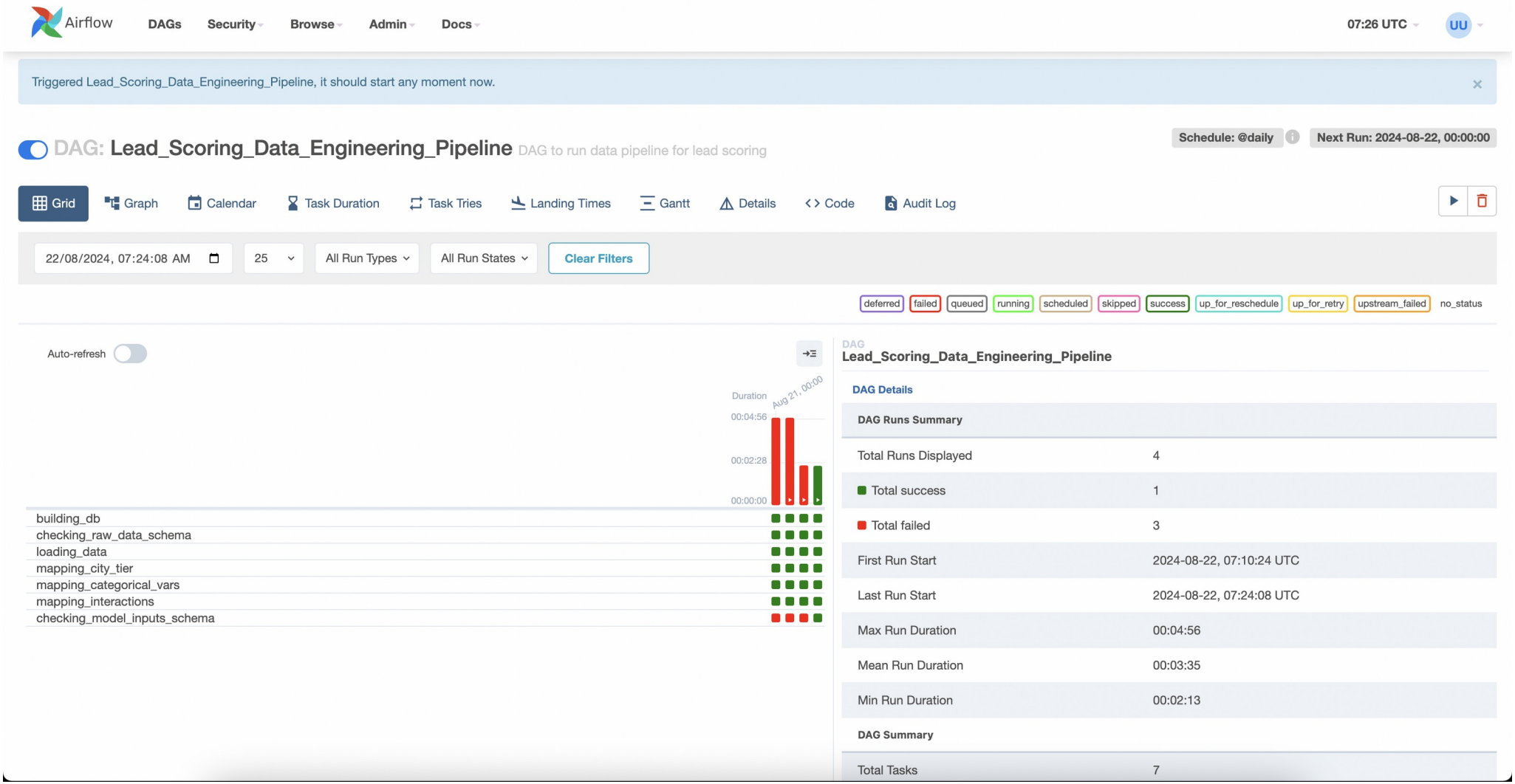
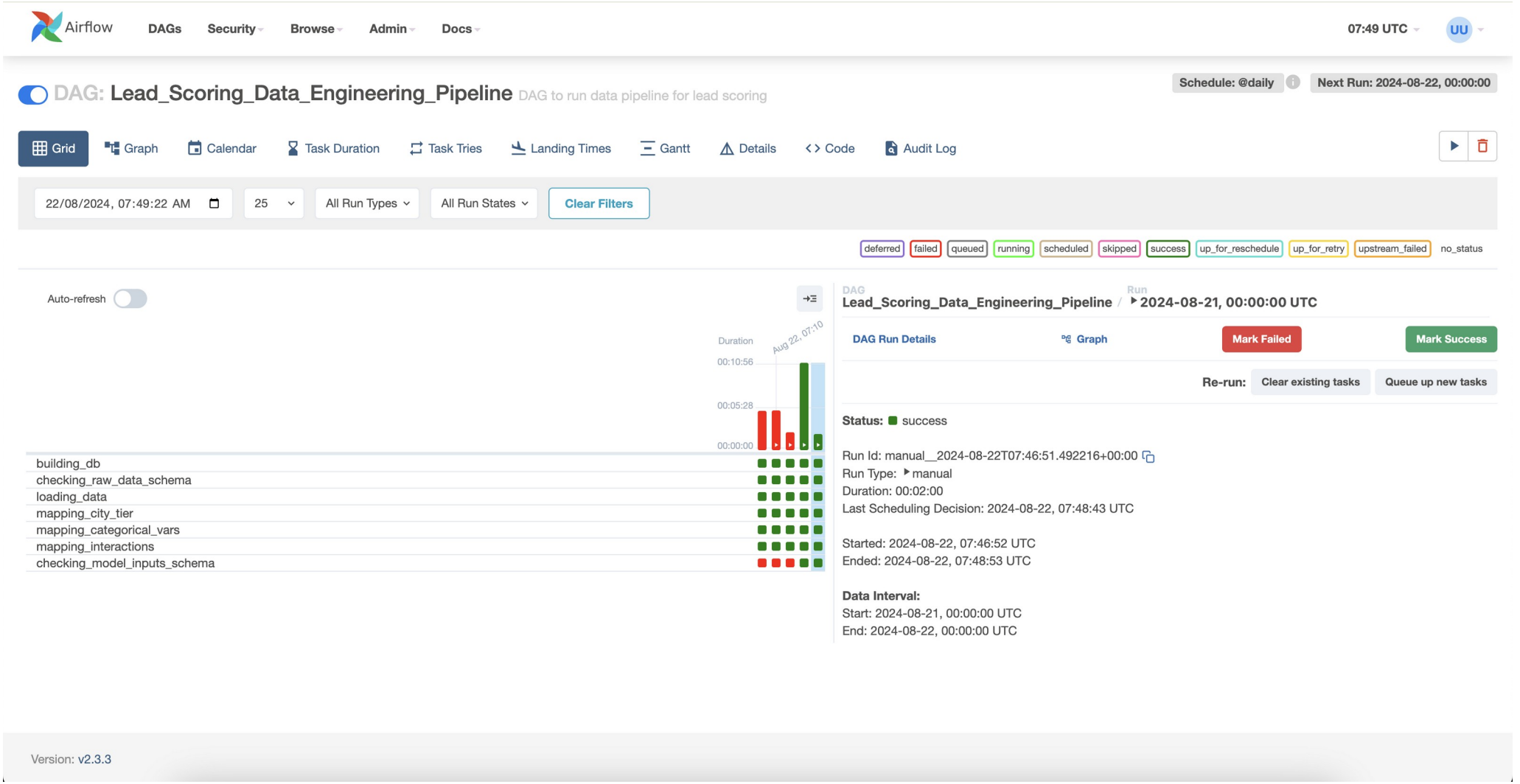
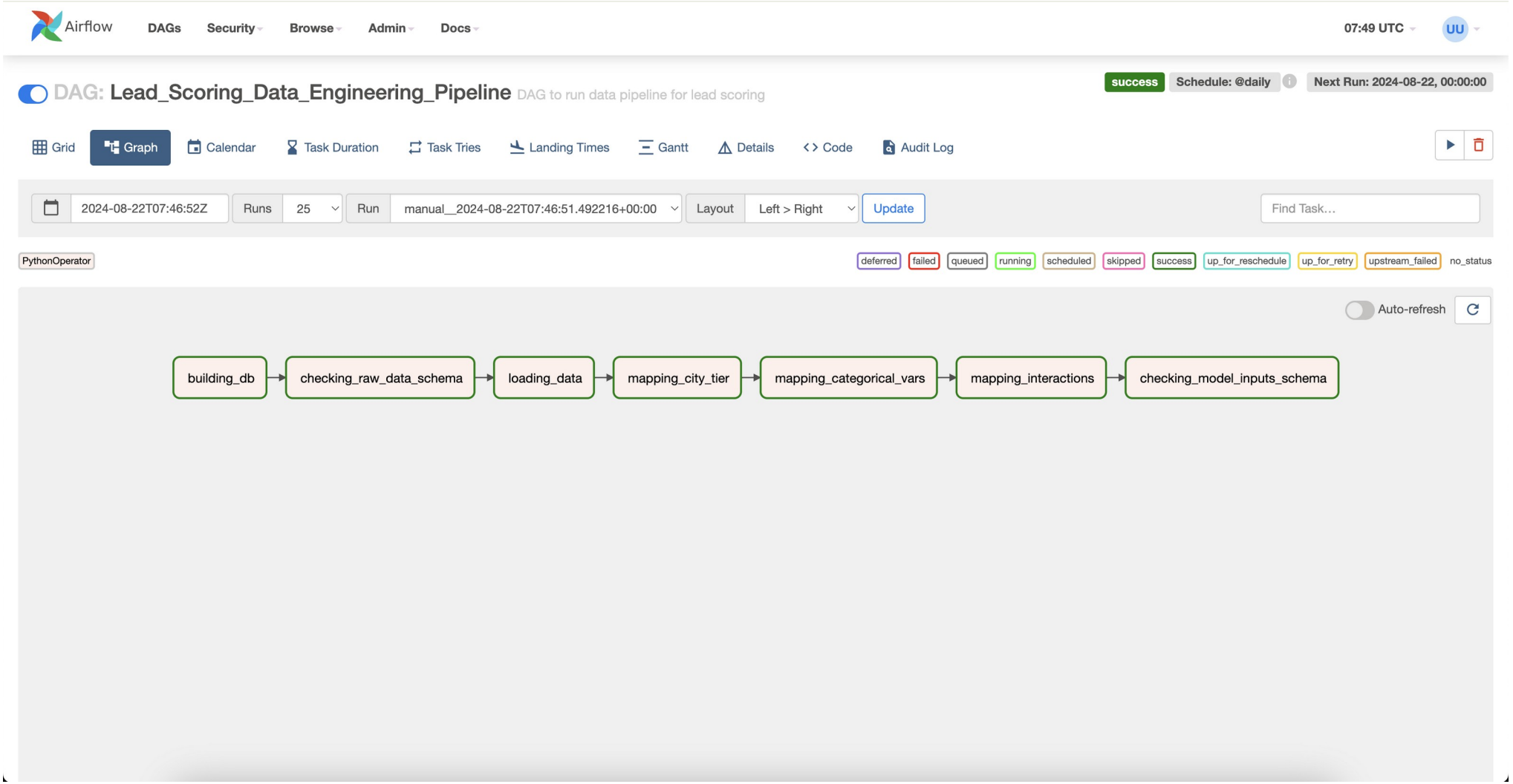
Data Pipeline:



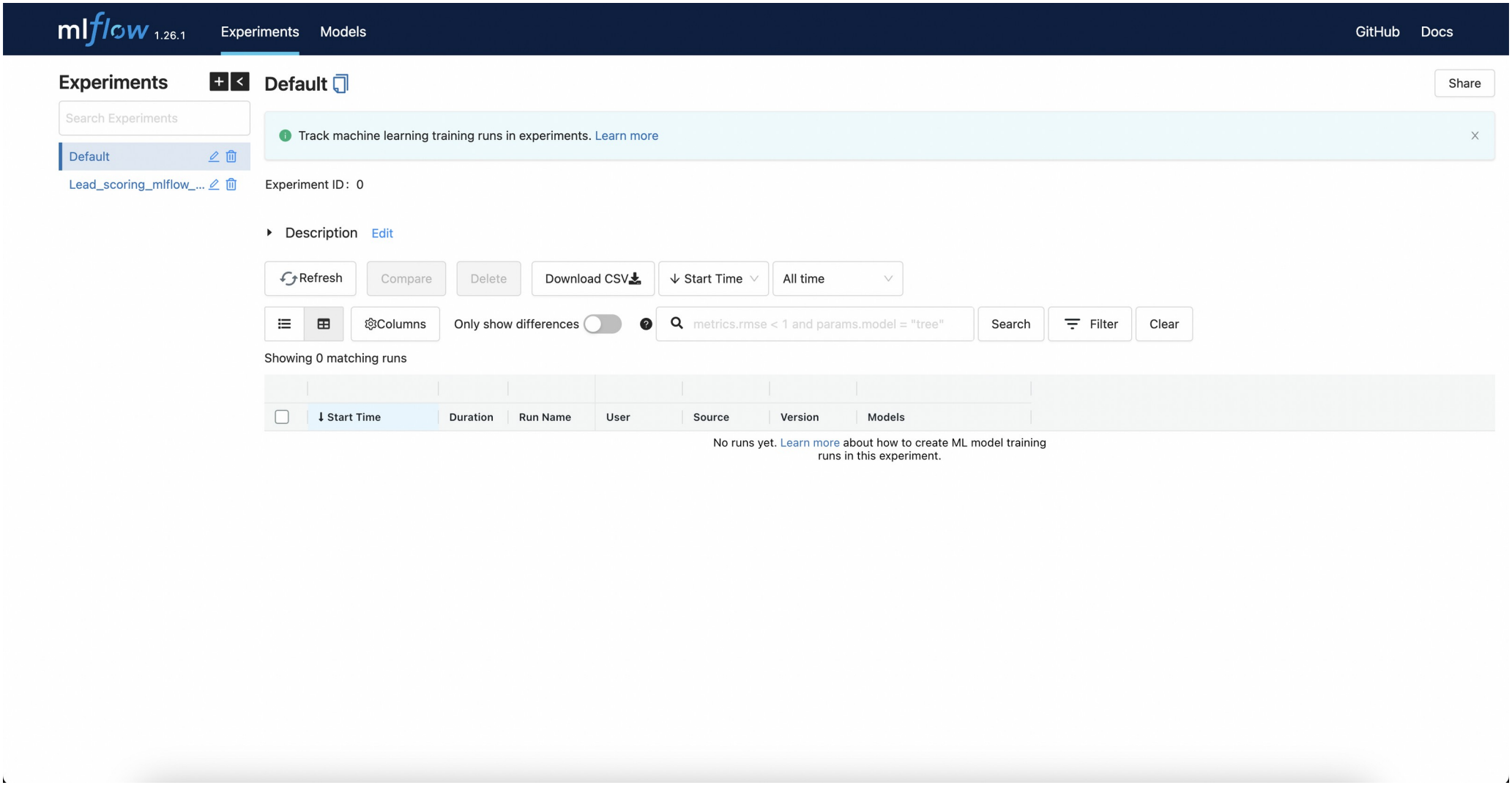

Airflow
DAGs
Security
Browse
Admin
Docs
07:10 UTC
UU
DAG: Lead_Scoring_Data_Engineering_Pipeline
DAG to run data pipeline for lead scoring
Schedule: @daily
Next Run: 2024-08-21, 00:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
22/08/2024, 07:10:03 AM
25
All Run Types
All Run States
Clear Filters
deferred
failed
queued
running
scheduled
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Auto-refresh
building_db
checking_raw_data_schema
loading_data
mapping_city_tier
mapping_categorical_vars
mapping_interactions
checking_model_inputs_schema
DAG
Lead_Scoring_Data_Engineering_Pipeline
DAG Details
DAG Summary
Total Tasks
7
PythonOperators
7
Version: v2.3.3

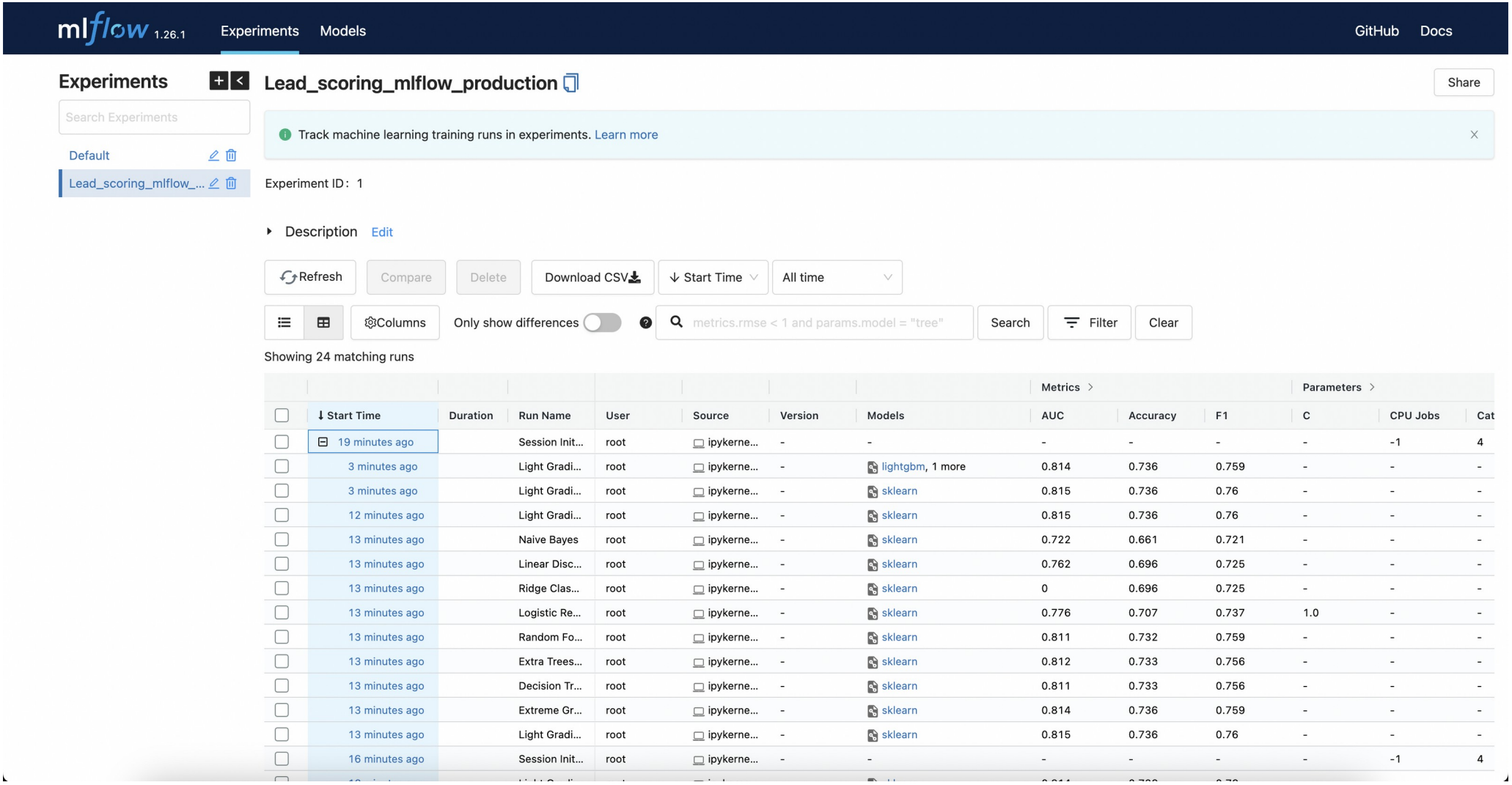

Airflow
DAGs
Security
Browse
Admin
Docs
07:26 UTC
UU
Triggered Lead_Scoring_Data_Engineering_Pipeline, it should start any moment now.
DAG: Lead_Scoring_Data_Engineering_Pipeline
DAG to run data pipeline for lead scoring
Schedule: @daily
Next Run: 2024-08-22, 00:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
22/08/2024, 07:24:08 AM
25
All Run Types
All Run States
Clear Filters
deferred
failed
queued
running
scheduled
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Auto-refresh
Duration
Aug 21, 00:00
00:04:56
00:02:28
00:00:00
building_db
checking_raw_data_schema
loading_data
mapping_city_tier
mapping_categorical_vars
mapping_interactions
checking_model_inputs_schema
DAG
Lead_Scoring_Data_Engineering_Pipeline
DAG Details
DAG Runs Summary
Total Runs Displayed 4
Total success 1
Total failed 3
First Run Start 2024-08-22, 07:10:24 UTC
Last Run Start 2024-08-22, 07:24:08 UTC
Max Run Duration 00:04:56
Mean Run Duration 00:03:35
Min Run Duration 00:02:13
DAG Summary
Total Tasks 7

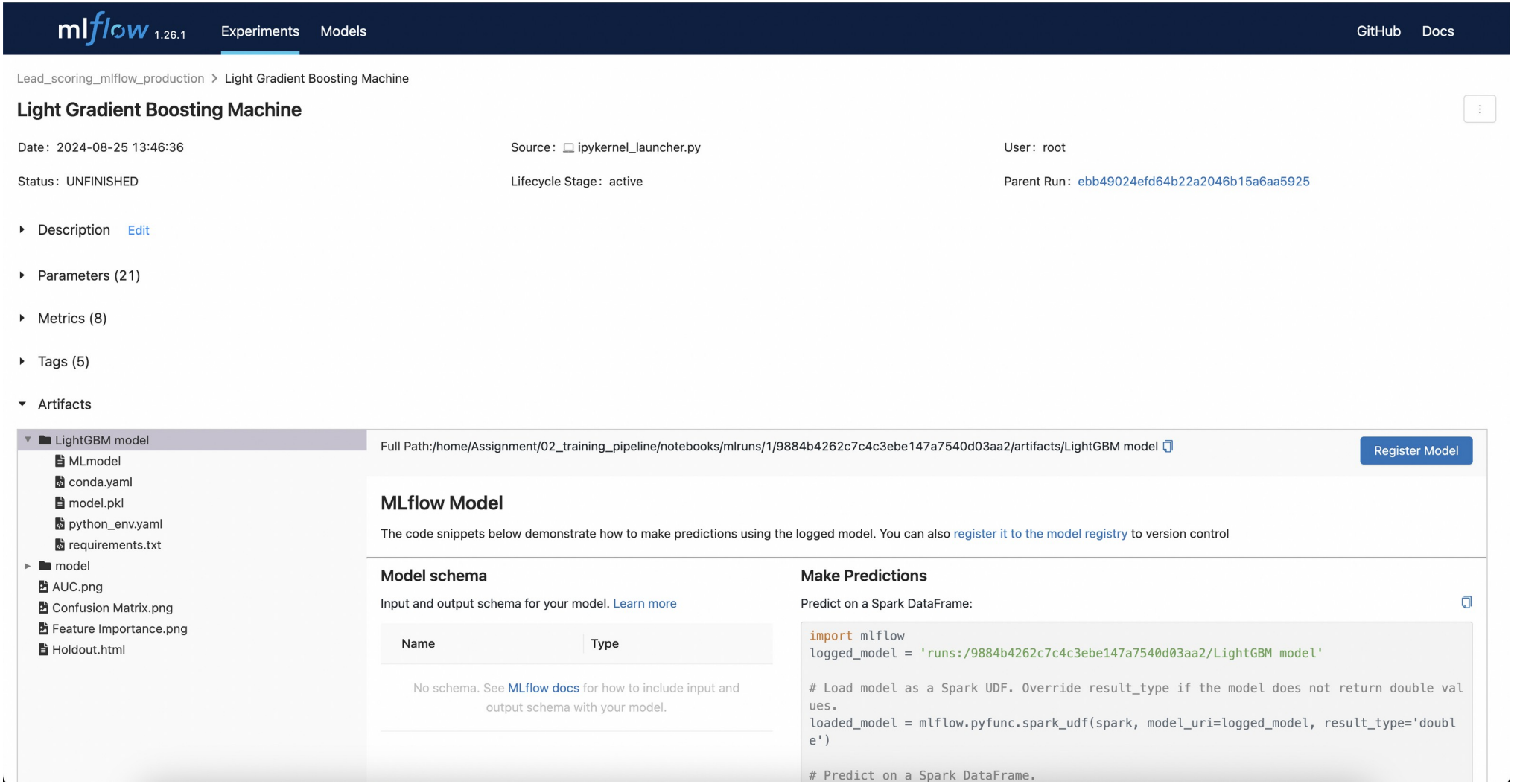

Airflow
DAGs
Security
Browse
Admin
Docs
07:49 UTC
UU
DAG: Lead_Scoring_Data_Engineering_Pipeline
DAG to run data pipeline for lead scoring
Schedule: @daily
Next Run: 2024-08-22, 00:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
22/08/2024, 07:49:22 AM
25
All Run Types
All Run States
Clear Filters
deferred
failed
queued
running
scheduled
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Auto-refresh
Duration
Aug 22, 07:10
00:10:56
00:05:28
00:00:00
building_db
checking_raw_data_schema
loading_data
mapping_city_tier
mapping_categorical_vars
mapping_interactions
checking_model_inputs_schema
DAG
Lead_Scoring_Data_Engineering_Pipeline
Run
2024-08-21, 00:00:00 UTC
DAG Run Details
Graph
Mark Failed
Mark Success
Re-run:
Clear existing tasks
Queue up new tasks
Status: success
Run Id: manual__2024-08-22T07:46:51.492216+00:00
Run Type: manual
Duration: 00:02:00
Last Scheduling Decision: 2024-08-22, 07:48:43 UTC
Started: 2024-08-22, 07:46:52 UTC
Ended: 2024-08-22, 07:48:53 UTC
Data Interval:
Start: 2024-08-21, 00:00:00 UTC
End: 2024-08-22, 00:00:00 UTC
Version: v2.3.3


Airflow
DAGs
Security
Browse
Admin
Docs
07:49 UTC
UU
DAG: Lead_Scoring_Data_Engineering_Pipeline
DAG to run data pipeline for lead scoring
success
Schedule: @daily
Next Run: 2024-08-22, 00:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
2024-08-22T07:46:52Z
Runs
25
Run
manual__2024-08-22T07:46:51.492216+00:00
Layout
Left > Right
Update
Find Task...
PythonOperator
deferred
failed
queued
running
scheduled
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Auto-refresh
building_db
checking_raw_data_schema
loading_data
mapping_city_tier
mapping_categorical_vars
mapping_interactions
checking_model_inputs_schema

Experiment on the cleaned data




mlflow 1.26.1
Experiments
Models
GitHub
Docs
Experiments
Search Experiments
Default
Lead_scoring_mlflow_...
Lead_scoring_mlflow_production
Share
Track machine learning training runs in experiments. Learn more
Experiment ID: 1
Description
Edit
Refresh
Compare
Delete
Download CSV
Start Time
All time
Columns
Only show differences
metrics.rmse < 1 and params.model = "tree"
Search
Filter
Clear
Showing 24 matching runs
Metrics
Parameters
Start Time
Duration
Run Name
User
Source
Version
Models
AUC
Accuracy
F1
C
CPU Jobs
Cat
19 minutes ago Session Init... root ipykerne... - - - - - - -1 4
3 minutes ago Light Gradi... root ipykerne... - lightgbm, 1 more 0.814 0.736 0.759 - - -
3 minutes ago Light Gradi... root ipykerne... - sklearn 0.815 0.736 0.76 - - -
12 minutes ago Light Gradi... root ipykerne... - sklearn 0.815 0.736 0.76 - - -
13 minutes ago Naive Bayes root ipykerne... - sklearn 0.722 0.661 0.721 - - -
13 minutes ago Linear Disc... root ipykerne... - sklearn 0.762 0.696 0.725 - - -
13 minutes ago Ridge Clas... root ipykerne... - sklearn 0 0.696 0.725 - - -
13 minutes ago Logistic Re... root ipykerne... - sklearn 0.776 0.707 0.737 1.0 - -
13 minutes ago Random Fo... root ipykerne... - sklearn 0.811 0.732 0.759 - - -
13 minutes ago Extra Trees... root ipykerne... - sklearn 0.812 0.733 0.756 - - -
13 minutes ago Decision Tr... root ipykerne... - sklearn 0.811 0.733 0.756 - - -
13 minutes ago Extreme Gr... root ipykerne... - sklearn 0.814 0.736 0.759 - - -
13 minutes ago Light Gradi... root ipykerne... - sklearn 0.815 0.736 0.76 - - -
16 minutes ago Session Init... root ipykerne... - - - - - - -1 4


mlflow 1.26.1
Experiments
Models
GitHub
Docs
Lead_scoring_mlflow_production > Light Gradient Boosting Machine
Light Gradient Boosting Machine
Date: 2024-08-25 13:46:36
Source: ipykernel_launcher.py
User: root
Status: UNFINISHED
Lifecycle Stage: active
Parent Run: ebb49024efd64b22a2046b15a6aa5925
Description
Edit
Parameters (21)
Metrics (8)
Tags (5)
Artifacts
LightGBM model
MLmodel
conda.yaml
model.pkl
python_env.yaml
requirements.txt
model
AUC.png
Confusion Matrix.png
Feature Importance.png
Holdout.html
Full Path:/home/Assignment/02_training_pipeline/notebooks/mlruns/1/9884b4262c7c4c3ebe147a7540d03aa2/artifacts/LightGBM model
Register Model
MLflow Model
The code snippets below demonstrate how to make predictions using the logged model. You can also register it to the model registry to version control
Model schema
Input and output schema for your model. Learn more
Name
Type
No schema. See MLflow docs for how to include input and output schema with your model.
Make Predictions
Predict on a Spark DataFrame:
import mlflow
logged_model = 'runs:/9884b4262c7c4c3ebe147a7540d03aa2/LightGBM model'
# Load model as a Spark UDF. Override result_type if the model does not return double values.
loaded_model = mlflow.pyfunc.spark_udf(spark, model_uri=logged_model, result_type='double')
# Predict on a Spark DataFrame.

▼ Parameters (21)

Name	Value
boosting_type	gbdt
class_weight	None
colsample_bytree	1.0
device	gpu
importance_type	split
learning_rate	0.1
max_depth	-1
min_child_samples	20
min_child_weight	0.001
min_split_gain	0.0
n_estimators	100
n_jobs	-1
num_leaves	41
objective	None
random_state	42
reg_alpha	0.0
reg_lambda	0.0

reg_alpha	0.0
reg_lambda	0.0
silent	warn
subsample	1.0
subsample_for_bin	200000
subsample_freq	0

▼ Metrics (8)

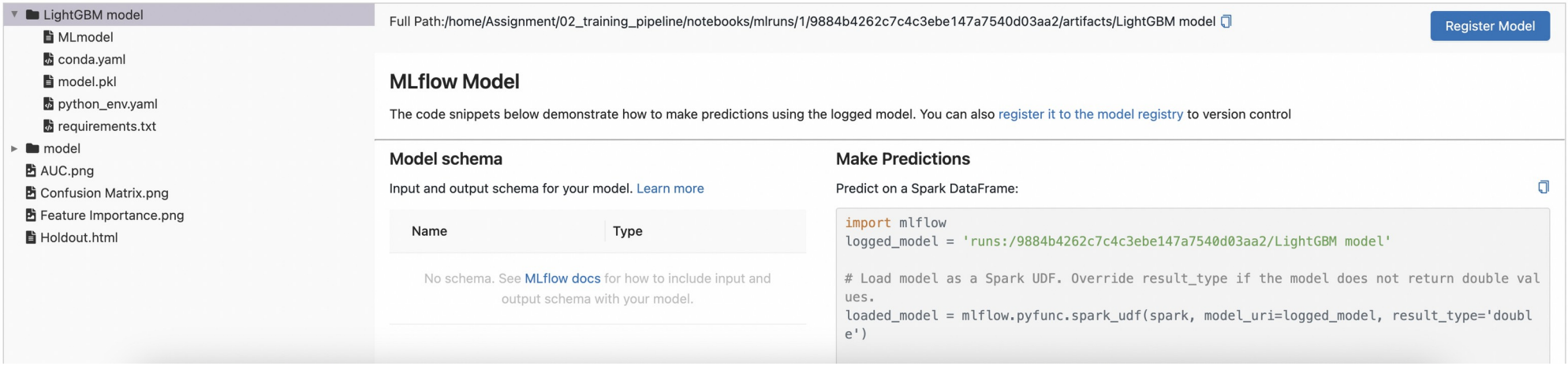
Name	Value
AUC	0.814
Accuracy	0.736
F1	0.759
Kappa	0.471
MCC	0.479
Prec.	0.699
Recall	0.832
TT	5.94

▼ Tags (5)

Name	Value	Actions
Run ID	9884b4262c7c4c3ebe147a7540d03aa2	
Run Time	51.06	
Source	finalize_model	
URI	df57f514	
USI	214c	

Name | Value | Add

▼ Artifacts



▸ Description Edit

▸ Parameters (21)

▸ Metrics (8)

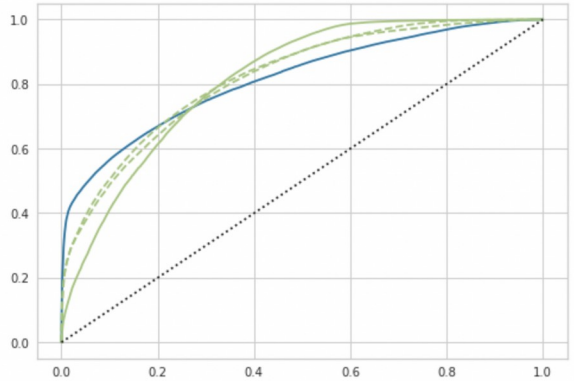
▸ Tags (5)

▾ Artifacts

- ▾ LightGBM model
  - MLmodel
  - conda.yaml
  - model.pkl
  - python_env.yaml
  - requirements.txt
- ▸ model
- AUC.png
- Confusion Matrix.png
- Feature Importance.png
- Holdout.html

Full Path:/home/Assignment/02_training_pipeline/notebooks/mlruns/1/9884b4262c7c4c3ebe147a7540d03aa2/artifacts/AUC.png

Size: 27.87KB



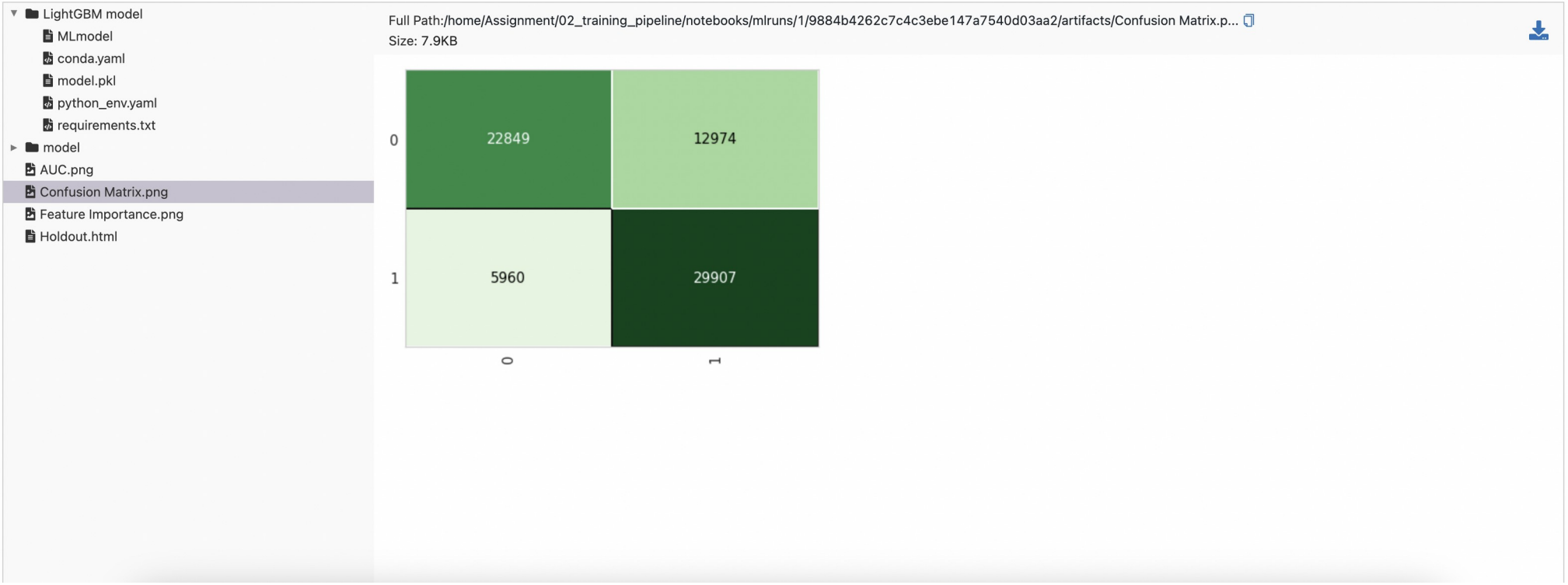
▸ Description Edit

▸ Parameters (21)

▸ Metrics (8)

▸ Tags (5)

▾ Artifacts



▸ Description Edit

▸ Parameters (21)

▸ Metrics (8)

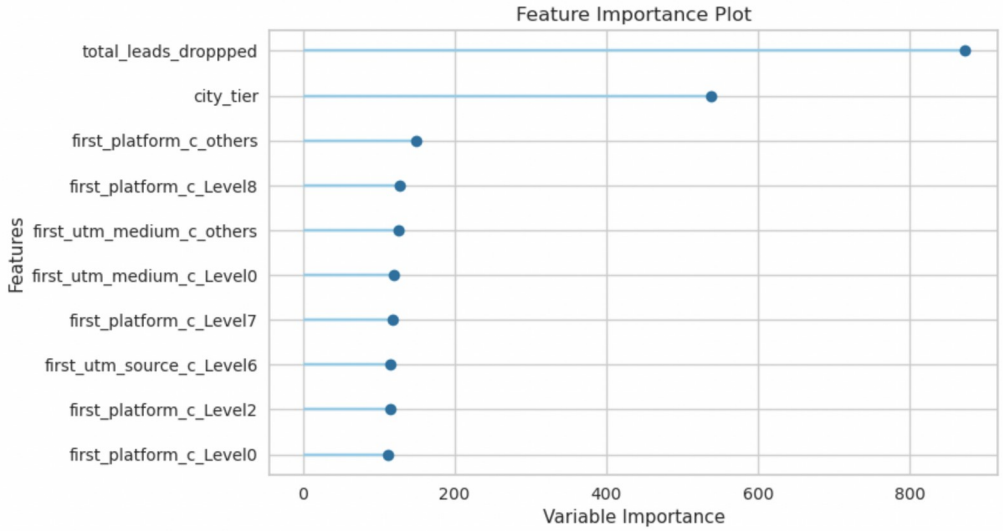
▸ Tags (5)

▾ Artifacts

- ▾ LightGBM model
  - MLmodel
  - conda.yaml
  - model.pkl
  - python_env.yaml
  - requirements.txt
- ▸ model
- AUC.png
- Confusion Matrix.png
- Feature Importance.png
- Holdout.html

Full Path:/home/Assignment/02_training_pipeline/notebooks/mlruns/1/9884b4262c7c4c3ebe147a7540d03aa2/artifacts/Feature Importanc...

Size: 41.03KB



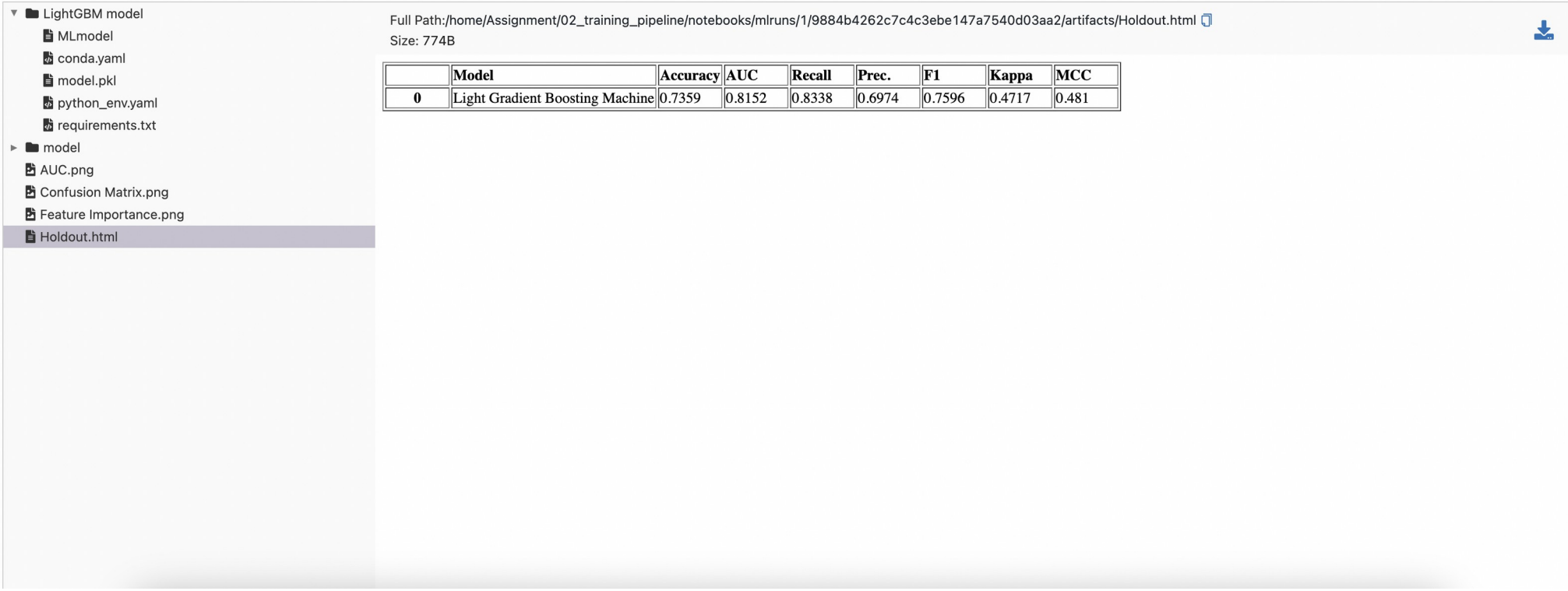
▸ Description Edit

▸ Parameters (21)

▸ Metrics (8)

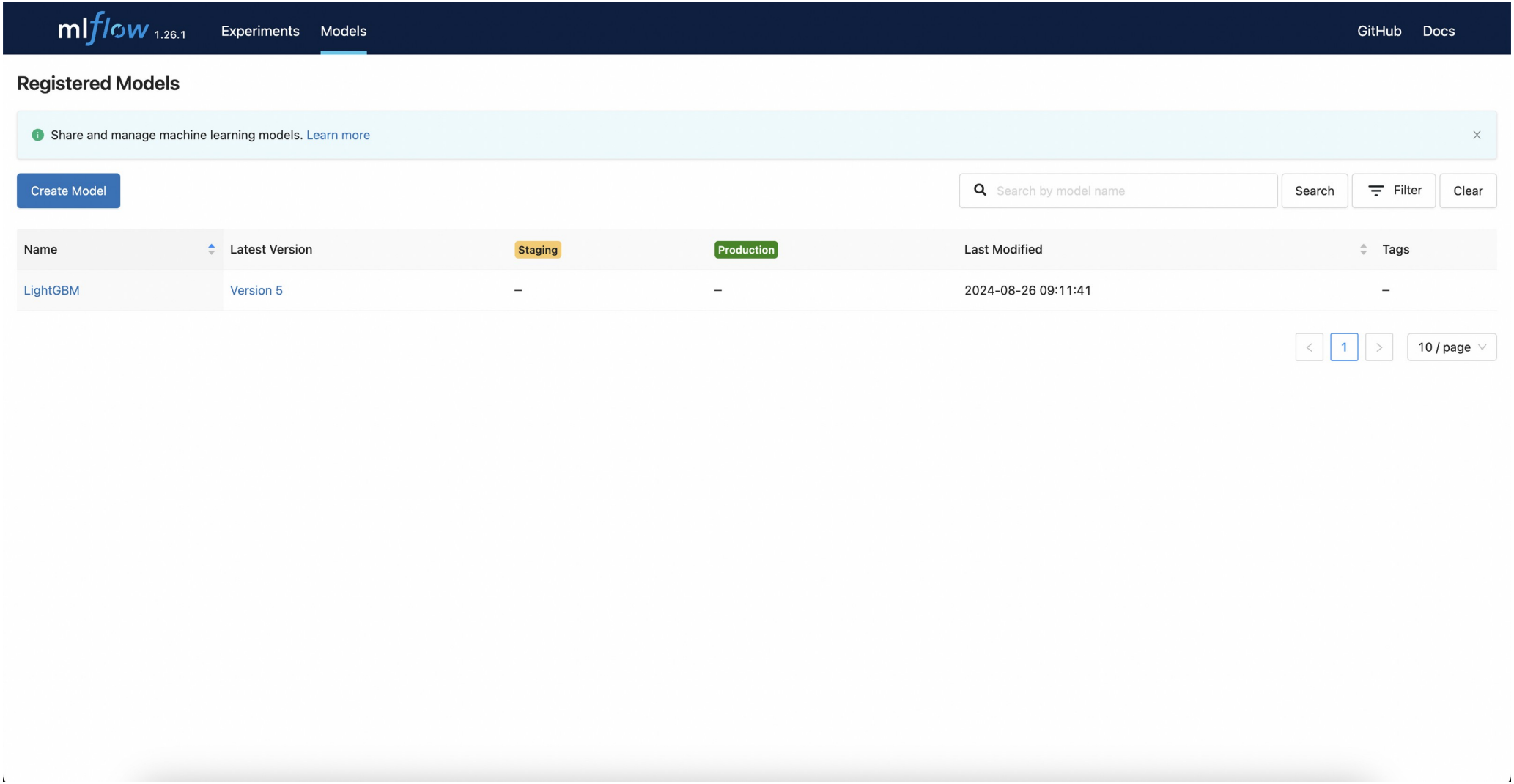
▸ Tags (5)

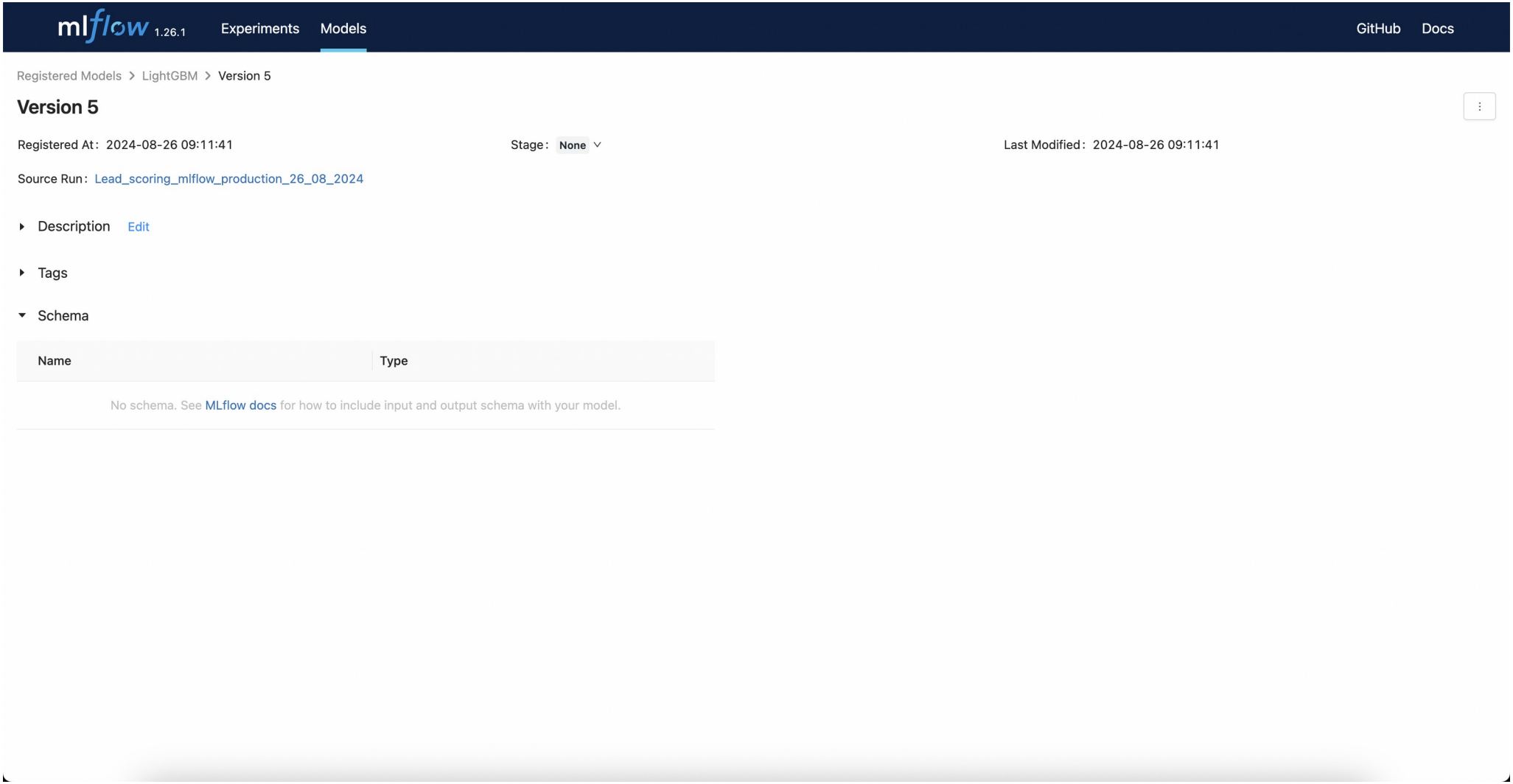
▾ Artifacts



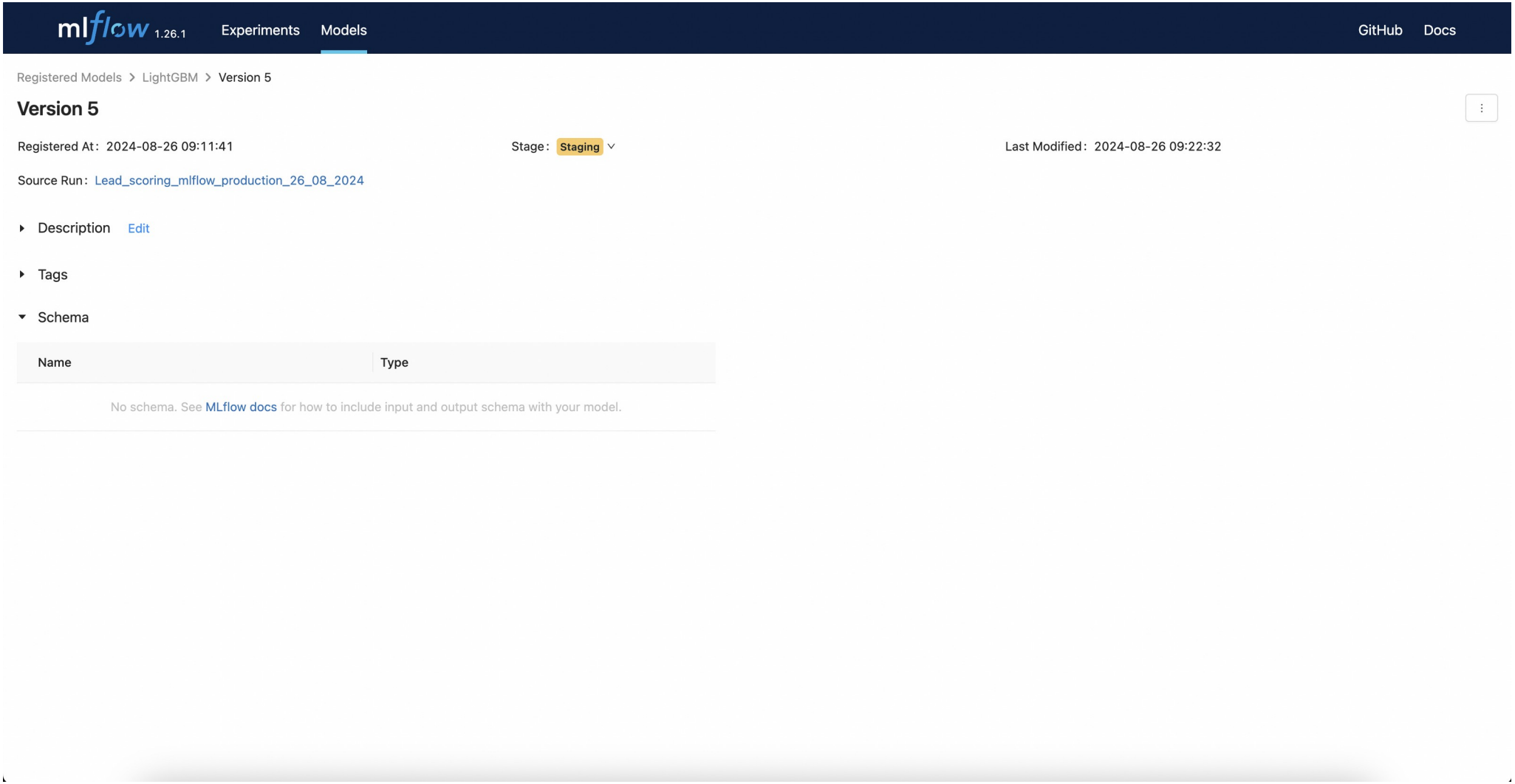
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Light Gradient Boosting Machine	0.7359	0.8152	0.8338	0.6974	0.7596	0.4717	0.481

Registering and Promoting model to Production

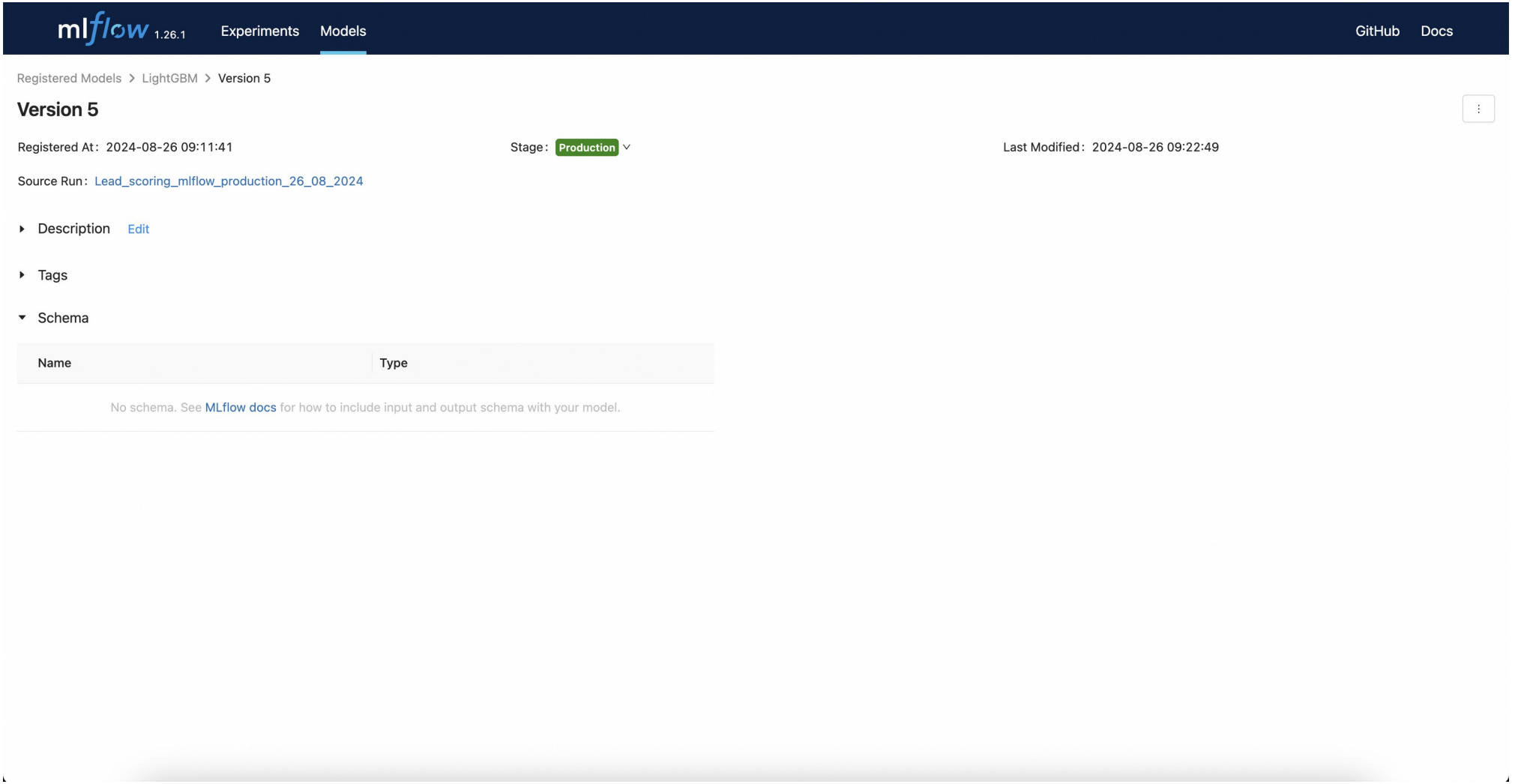


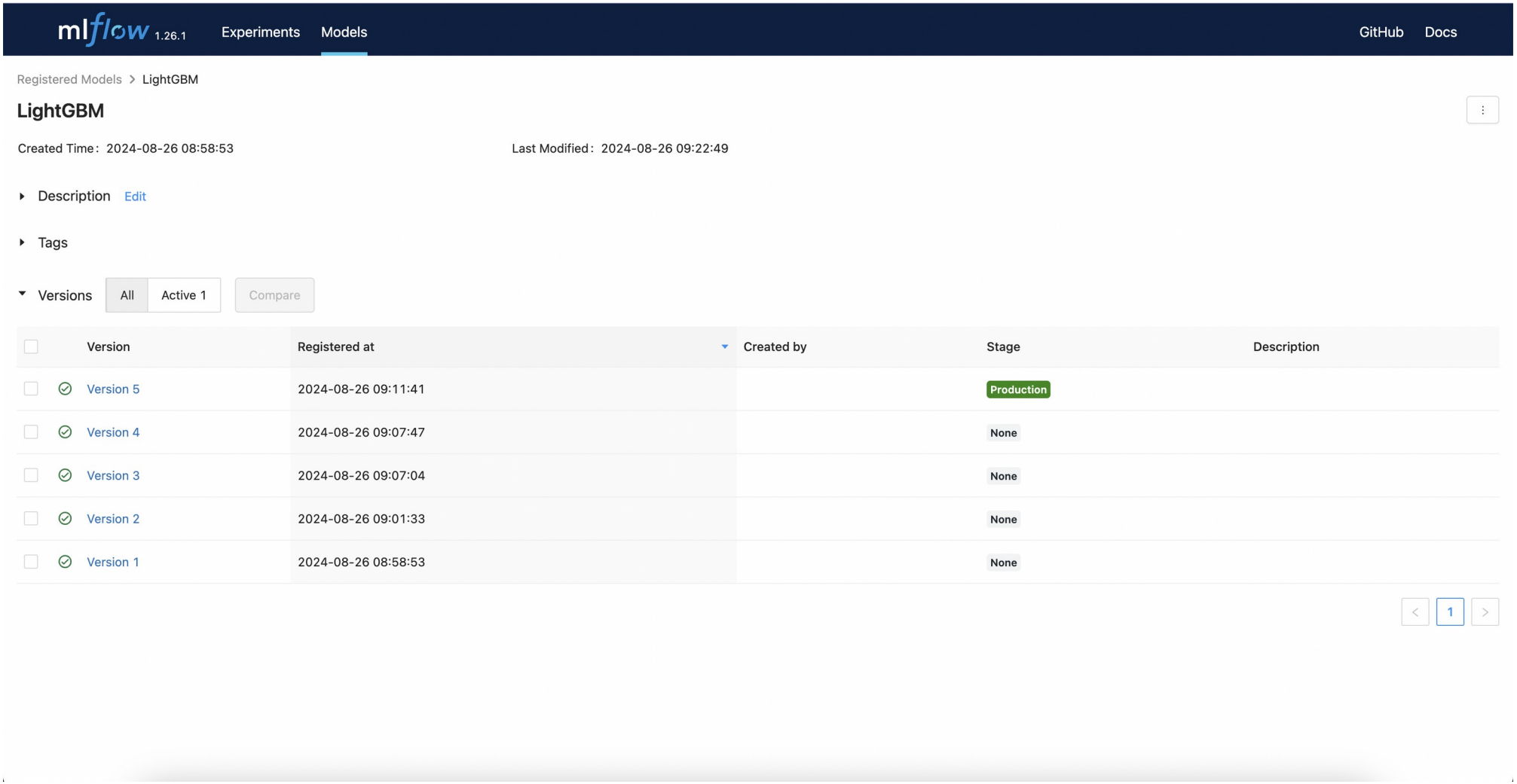

mlflow 1.26.1
Experiments
Models
GitHub
Docs
Registered Models > LightGBM > Version 5
Version 5
Registered At: 2024-08-26 09:11:41
Stage: None
Last Modified: 2024-08-26 09:11:41
Source Run: Lead_scoring_mlflow_production_26_08_2024
Description
Edit
Tags
Schema
Name
Type
No schema. See MLflow docs for how to include input and output schema with your model.

Promoting to Staging

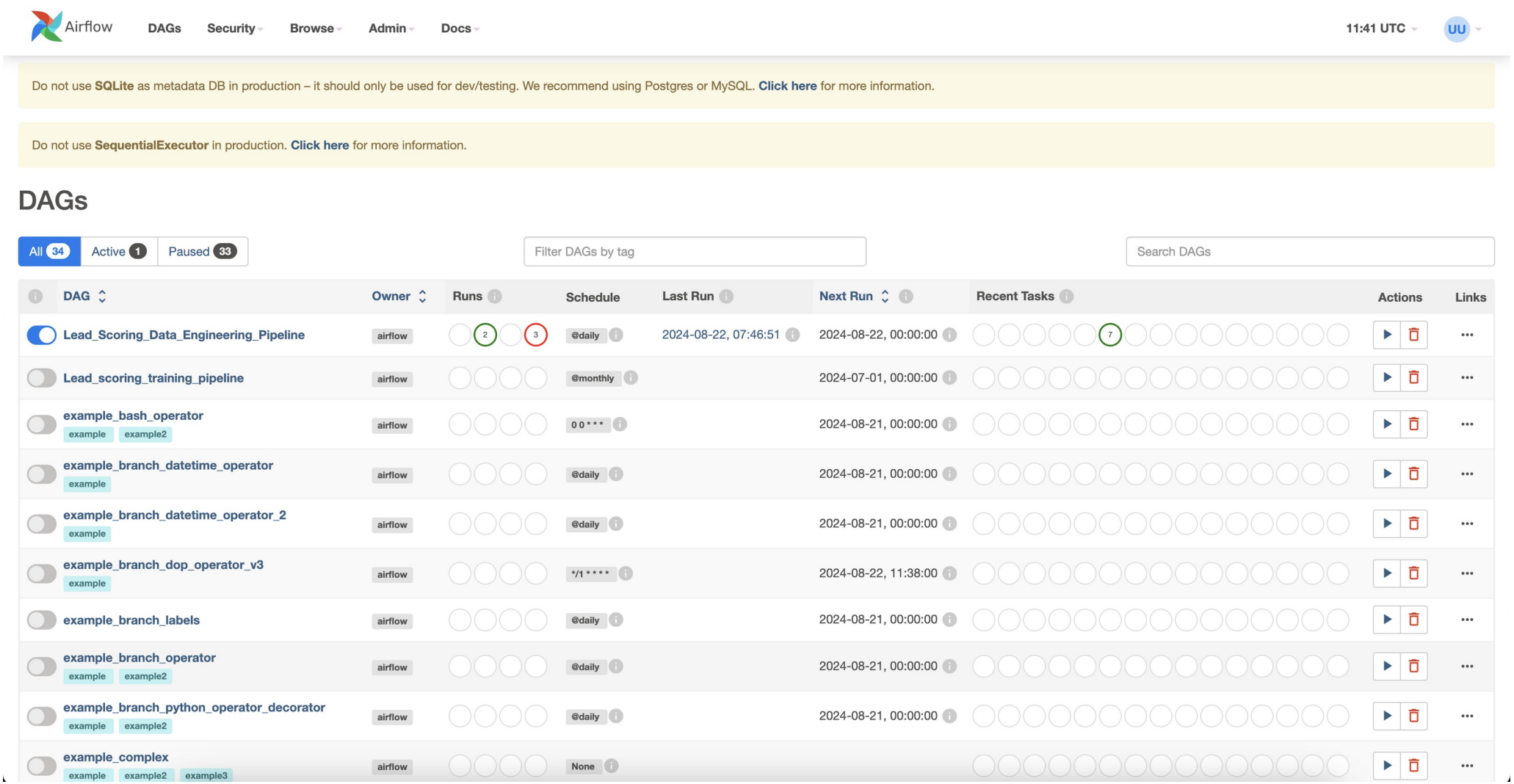


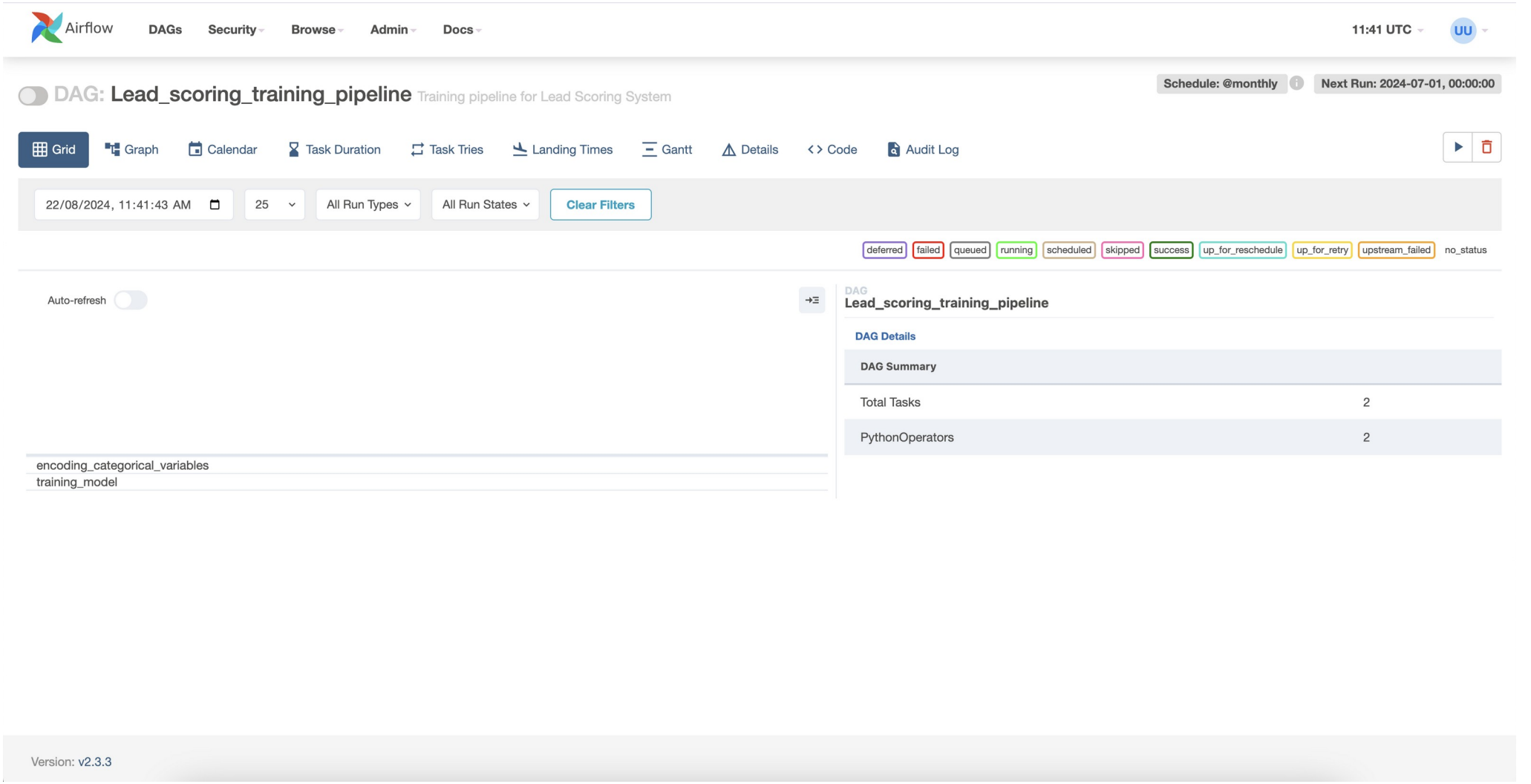
Promoting to Production

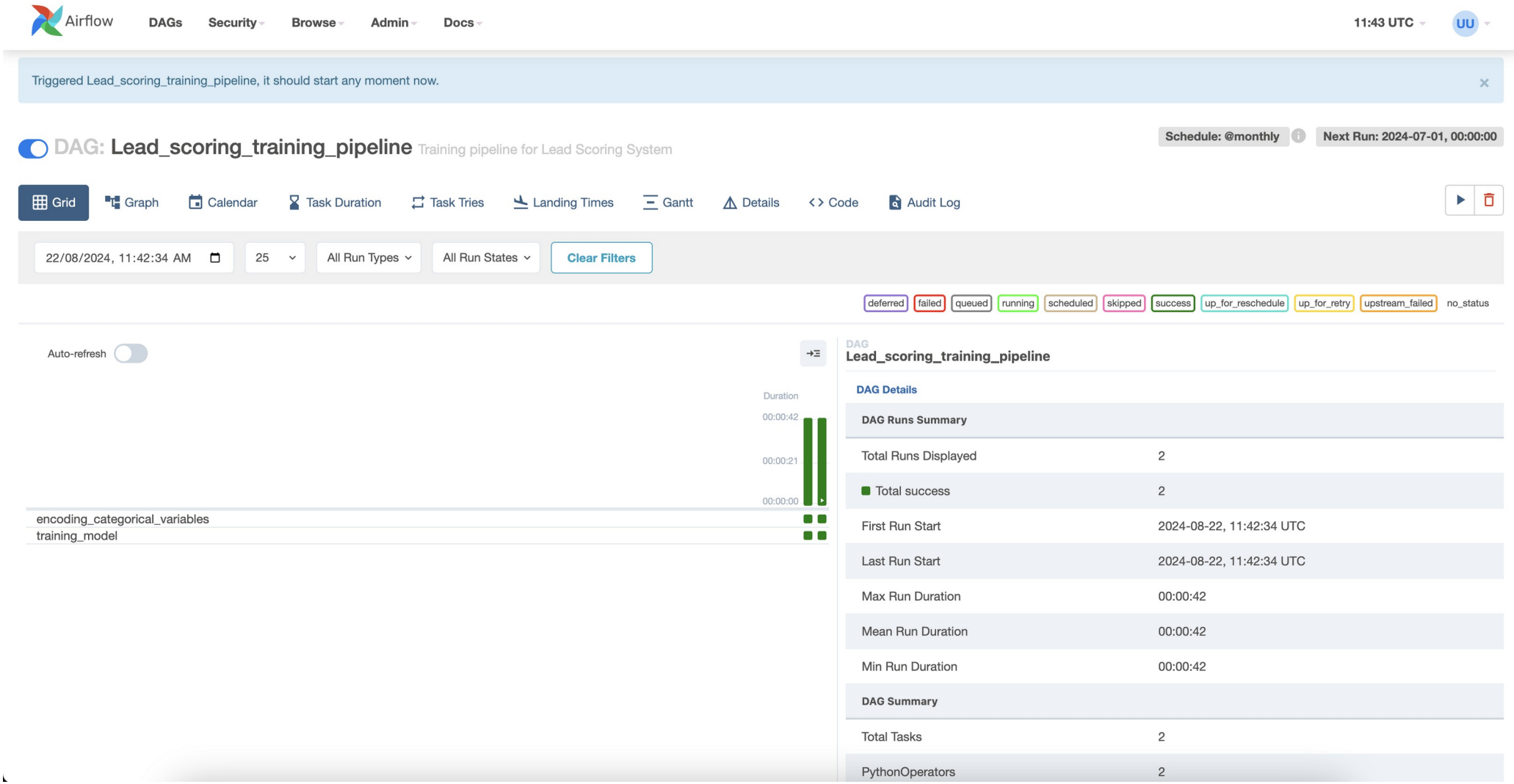


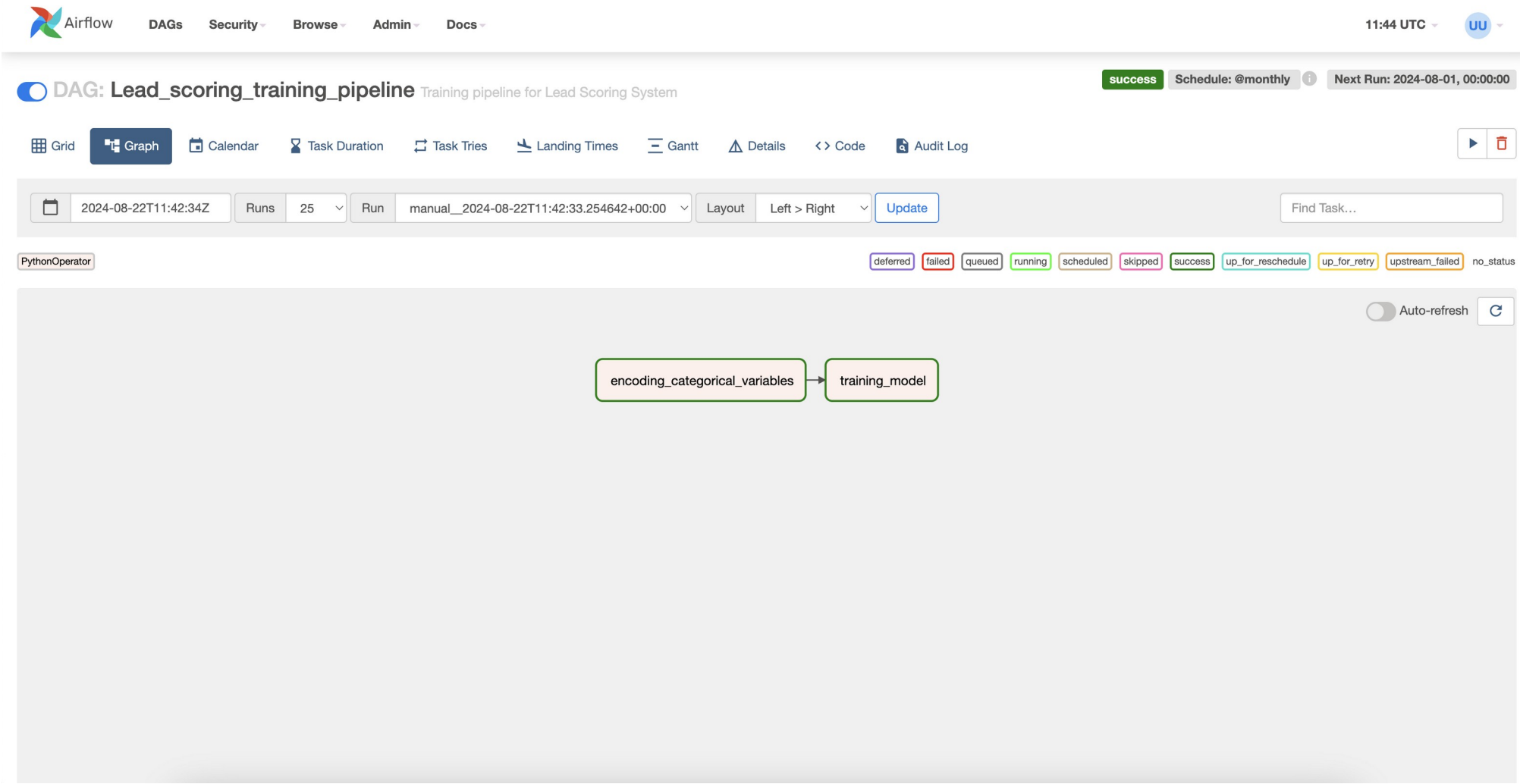

mlflow 1.26.1
Experiments
Models
GitHub
Docs
Registered Models > LightGBM
LightGBM
Created Time: 2024-08-26 08:58:53
Last Modified: 2024-08-26 09:22:49
Description
Edit
Tags
Versions
All
Active 1
Compare
Version
Registered at
Created by
Stage
Description
Version 5
2024-08-26 09:11:41
Production
Version 4
2024-08-26 09:07:47
None
Version 3
2024-08-26 09:07:04
None
Version 2
2024-08-26 09:01:33
None
Version 1
2024-08-26 08:58:53
None
1

Training Pipeline

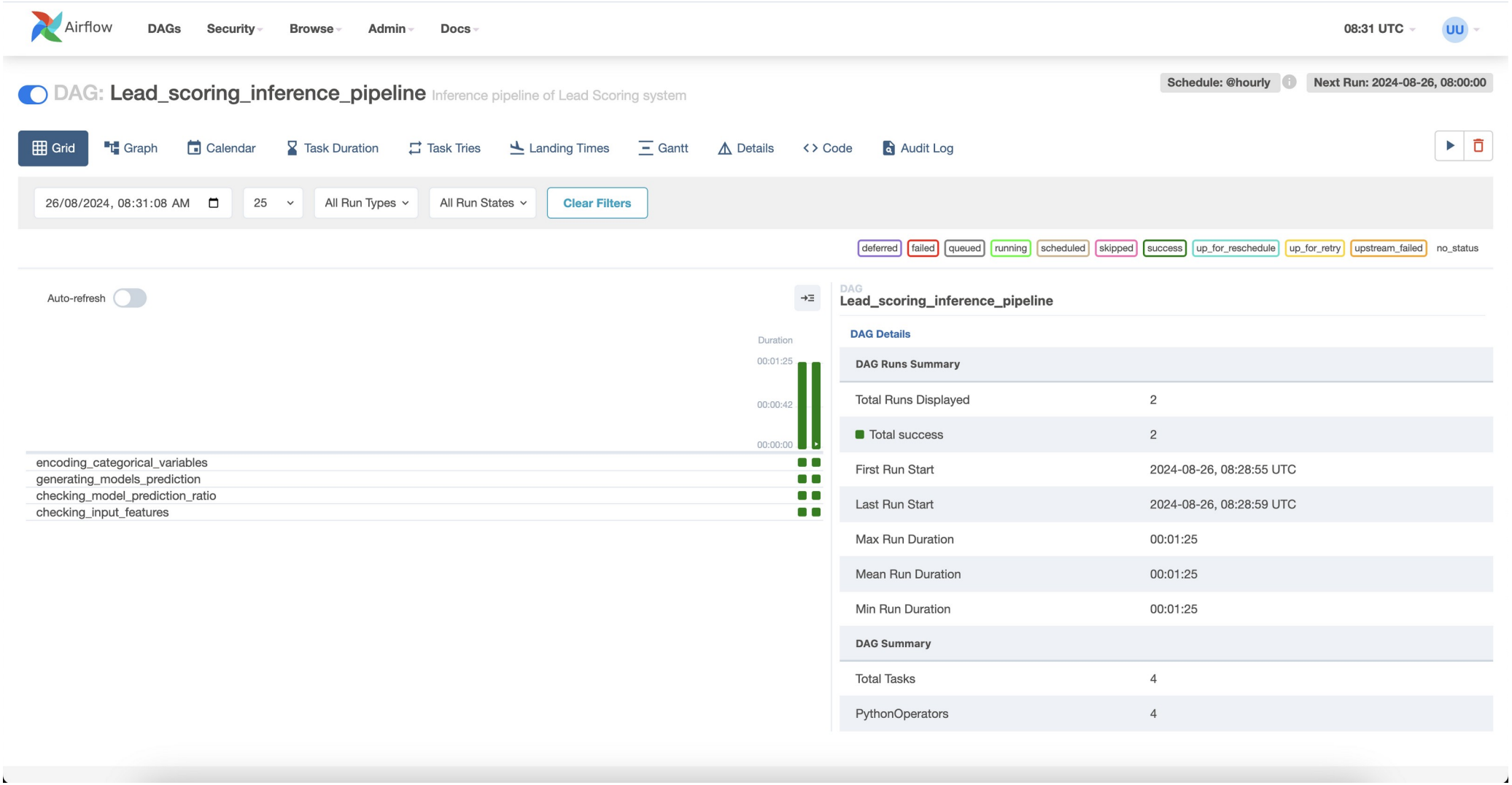


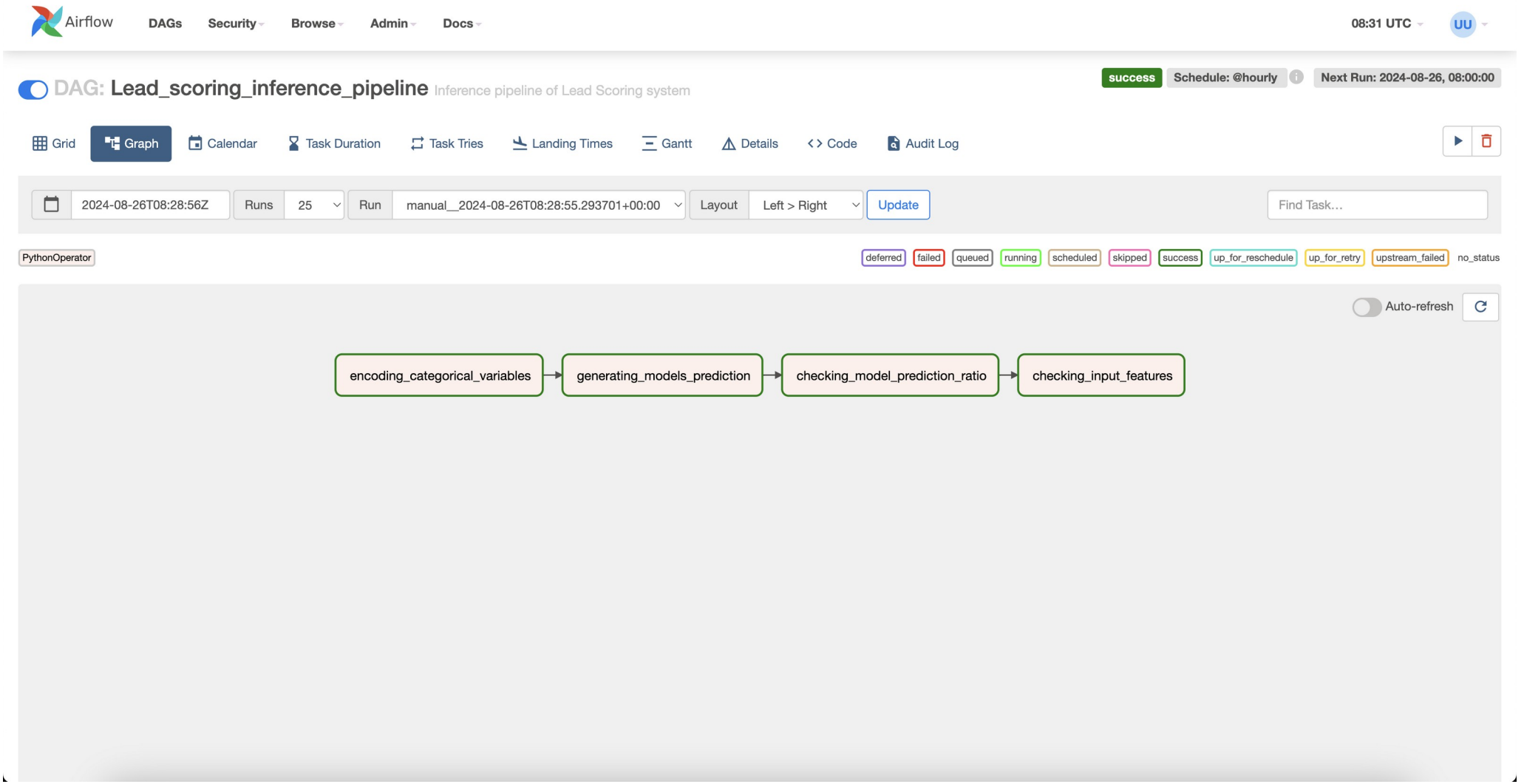

Airflow
DAGs
Security
Browse
Admin
Docs
11:41 UTC
UU
DAG: Lead_scoring_training_pipeline
Training pipeline for Lead Scoring System
Schedule: @monthly
Next Run: 2024-07-01, 00:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
22/08/2024, 11:41:43 AM
25
All Run Types
All Run States
Clear Filters
deferred
failed
queued
running
scheduled
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Auto-refresh
encoding_categorical_variables
training_model
DAG
Lead_scoring_training_pipeline
DAG Details
DAG Summary
Total Tasks
2
PythonOperators
2
Version: v2.3.3


Airflow
DAGs
Security
Browse
Admin
Docs
11:43 UTC
UU
Triggered Lead_scoring_training_pipeline, it should start any moment now.
DAG: Lead_scoring_training_pipeline
Training pipeline for Lead Scoring System
Schedule: @monthly
Next Run: 2024-07-01, 00:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
22/08/2024, 11:42:34 AM
25
All Run Types
All Run States
Clear Filters
deferred
failed
queued
running
scheduled
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Auto-refresh
Duration
00:00:42
00:00:21
00:00:00
encoding_categorical_variables
training_model
DAG
Lead_scoring_training_pipeline
DAG Details
DAG Runs Summary
Total Runs Displayed 2
Total success 2
First Run Start 2024-08-22, 11:42:34 UTC
Last Run Start 2024-08-22, 11:42:34 UTC
Max Run Duration 00:00:42
Mean Run Duration 00:00:42
Min Run Duration 00:00:42
DAG Summary
Total Tasks 2
PythonOperators 2


Airflow
DAGs
Security
Browse
Admin
Docs
11:44 UTC
UU
DAG: Lead_scoring_training_pipeline
Training pipeline for Lead Scoring System
success
Schedule: @monthly
Next Run: 2024-08-01, 00:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
2024-08-22T11:42:34Z
Runs
25
Run
manual__2024-08-22T11:42:33.254642+00:00
Layout
Left > Right
Update
Find Task...
PythonOperator
deferred
failed
queued
running
scheduled
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Auto-refresh
encoding_categorical_variables
training_model

Inference Pipeline




Airflow
DAGs
Security
Browse
Admin
Docs
08:31 UTC
UU
DAG: Lead_scoring_inference_pipeline
Inference pipeline of Lead Scoring system
success
Schedule: @hourly
Next Run: 2024-08-26, 08:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
2024-08-26T08:28:56Z
Runs
25
Run
manual__2024-08-26T08:28:55.293701+00:00
Layout
Left > Right
Update
Find Task...
PythonOperator
deferred
failed
queued
running
scheduled
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Auto-refresh
encoding_categorical_variables
generating_models_prediction
checking_model_prediction_ratio
checking_input_features