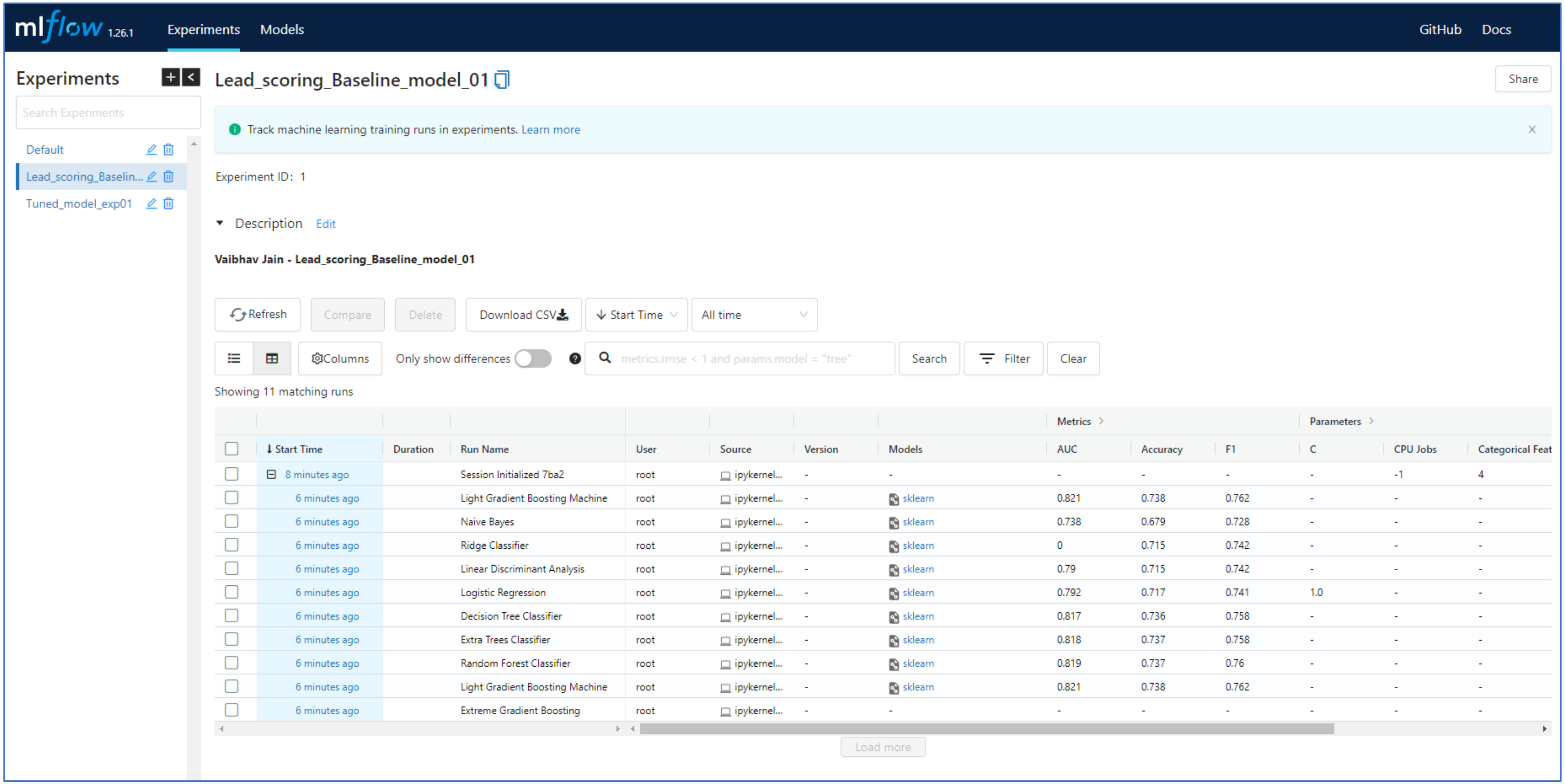
Assignment – CodePro Leadscoring MLOps – Vaibhav Jain

Model Experimentation - lead_scoring_model_experimentation.db

MLFlow UI Baseline model



Lead_scoring_Baseline_model_01

Track machine learning training runs in experiments. Learn more

Experiment ID: 1

Description Edit

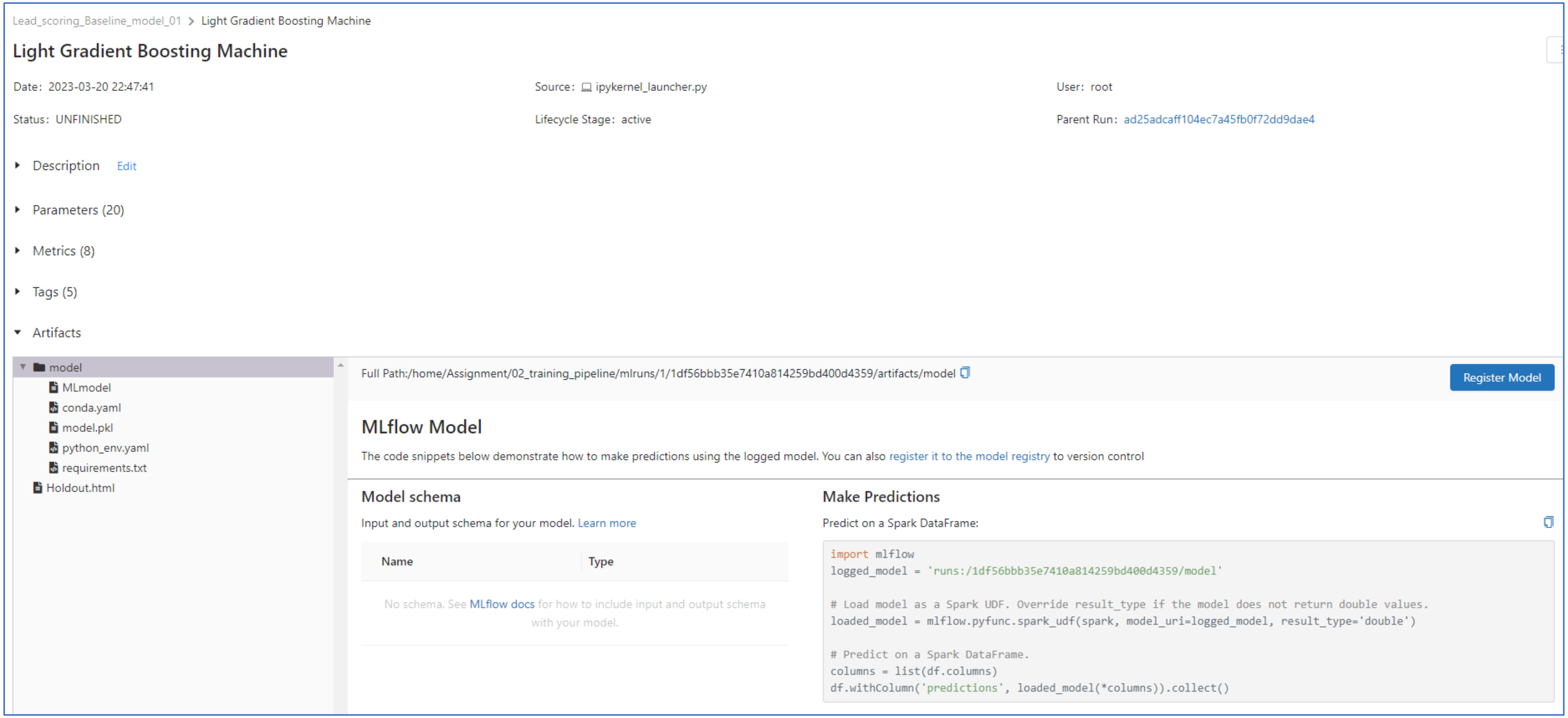
Vaibhav Jain - Lead_scoring_Baseline_model_01

metrics.rmse < 1 and params.model = "tree"

Showing 11 matching runs

□	Start Time	Duration	Run Name	User	Source	Version	Models	Metrics AUC	Accuracy	F1	Parameters C	CPU Jobs	Categorical Feat
□	⊟ 8 minutes ago		Session Initialized 7ba2	root	ipykernel...	-	-	-	-	-	-	-1	4
□	6 minutes ago		Light Gradient Boosting Machine	root	ipykernel...	-	sklearn	0.821	0.738	0.762	-	-	-
□	6 minutes ago		Naive Bayes	root	ipykernel...	-	sklearn	0.738	0.679	0.728	-	-	-
□	6 minutes ago		Ridge Classifier	root	ipykernel...	-	sklearn	0	0.715	0.742	-	-	-
□	6 minutes ago		Linear Discriminant Analysis	root	ipykernel...	-	sklearn	0.79	0.715	0.742	-	-	-
□	6 minutes ago		Logistic Regression	root	ipykernel...	-	sklearn	0.792	0.717	0.741	1.0	-	-
□	6 minutes ago		Decision Tree Classifier	root	ipykernel...	-	sklearn	0.817	0.736	0.758	-	-	-
□	6 minutes ago		Extra Trees Classifier	root	ipykernel...	-	sklearn	0.818	0.737	0.758	-	-	-
□	6 minutes ago		Random Forest Classifier	root	ipykernel...	-	sklearn	0.819	0.737	0.76	-	-	-
□	6 minutes ago		Light Gradient Boosting Machine	root	ipykernel...	-	sklearn	0.821	0.738	0.762	-	-	-
□	6 minutes ago		Extreme Gradient Boosting	root	ipykernel...	-	-	-	-	-	-	-	-

MLFlow UI Baseline model – One model with all artifacts



Lead_scoring_Baseline_model_01 > Light Gradient Boosting Machine

Light Gradient Boosting Machine

Date: 2023-03-20 22:47:41

Source: ipykernel_launcher.py

User: root

Status: UNFINISHED

Lifecycle Stage: active

Parent Run: ad25adcaff104ec7a45fb0f72dd9dae4

Full Path:/home/Assignment/02_training_pipeline/mlruns/1/1df56bbb35e7410a814259bd400d4359/artifacts/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

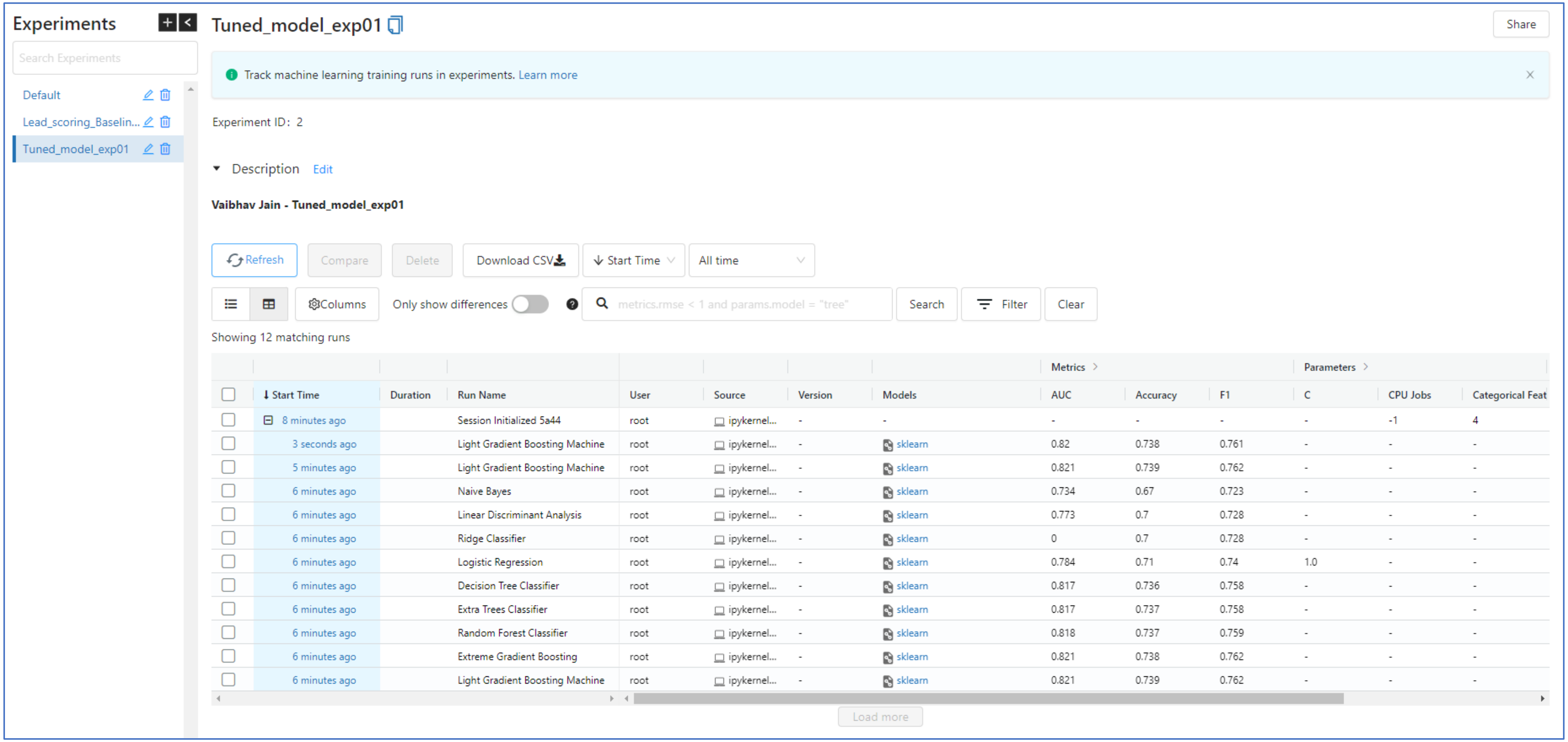
Predict on a Spark DataFrame:

```
import mlflow
logged_model = 'runs:/1df56bbb35e7410a814259bd400d4359/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.
columns = list(df.columns)
df.withColumn('predictions', loaded_model(*columns)).collect()
```

MLFlow UI Tuned model (Features dropped)



Tuned_model_exp01

Track machine learning training runs in experiments. Learn more

Experiment ID: 2

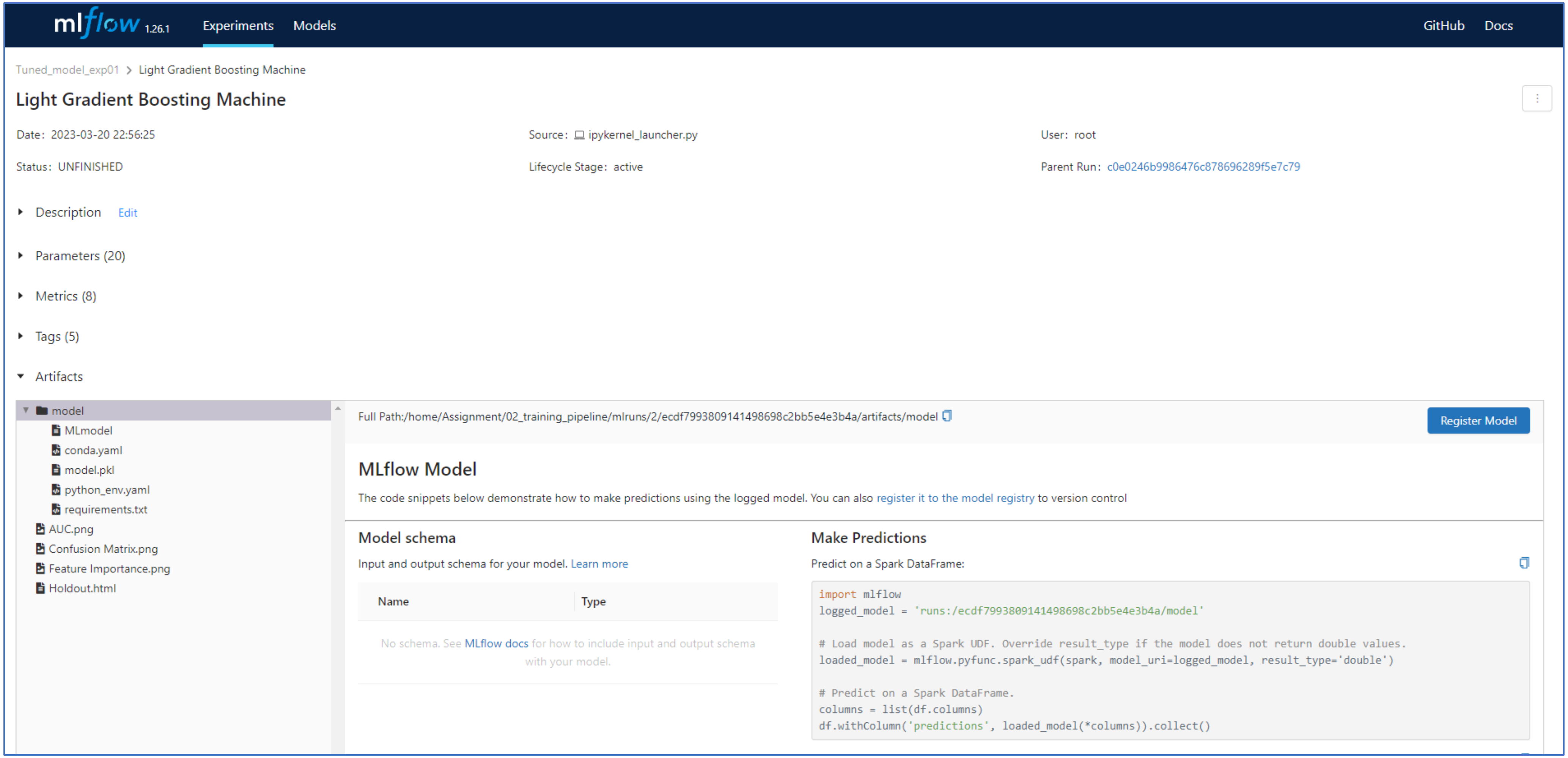
Description Edit

Vaibhav Jain - Tuned_model_exp01

Showing 12 matching runs

								Metrics			Parameters		
□	Start Time	Duration	Run Name	User	Source	Version	Models	AUC	Accuracy	F1	C	CPU Jobs	Categorical Feat
□	⊟ 8 minutes ago		Session Initialized 5a44	root	ipykernel...	-	-	-	-	-	-	-1	4
□	3 seconds ago		Light Gradient Boosting Machine	root	ipykernel...	-	sklearn	0.82	0.738	0.761	-	-	-
□	5 minutes ago		Light Gradient Boosting Machine	root	ipykernel...	-	sklearn	0.821	0.739	0.762	-	-	-
□	6 minutes ago		Naive Bayes	root	ipykernel...	-	sklearn	0.734	0.67	0.723	-	-	-
□	6 minutes ago		Linear Discriminant Analysis	root	ipykernel...	-	sklearn	0.773	0.7	0.728	-	-	-
□	6 minutes ago		Ridge Classifier	root	ipykernel...	-	sklearn	0	0.7	0.728	-	-	-
□	6 minutes ago		Logistic Regression	root	ipykernel...	-	sklearn	0.784	0.71	0.74	1.0	-	-
□	6 minutes ago		Decision Tree Classifier	root	ipykernel...	-	sklearn	0.817	0.736	0.758	-	-	-
□	6 minutes ago		Extra Trees Classifier	root	ipykernel...	-	sklearn	0.817	0.737	0.758	-	-	-
□	6 minutes ago		Random Forest Classifier	root	ipykernel...	-	sklearn	0.818	0.737	0.759	-	-	-
□	6 minutes ago		Extreme Gradient Boosting	root	ipykernel...	-	sklearn	0.821	0.738	0.762	-	-	-
□	6 minutes ago		Light Gradient Boosting Machine	root	ipykernel...	-	sklearn	0.821	0.739	0.762	-	-	-

MLFlow UI – Best Model with all artifacts



Tuned_model_exp01 > Light Gradient Boosting Machine

Light Gradient Boosting Machine

Date: 2023-03-20 22:56:25

Source: ipykernel_launcher.py

User: root

Status: UNFINISHED

Lifecycle Stage: active

Parent Run: c0e0246b9986476c878696289f5e7c79

Full Path:/home/Assignment/02_training_pipeline/mlruns/2/ecdf7993809141498698c2bb5e4e3b4a/artifacts/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

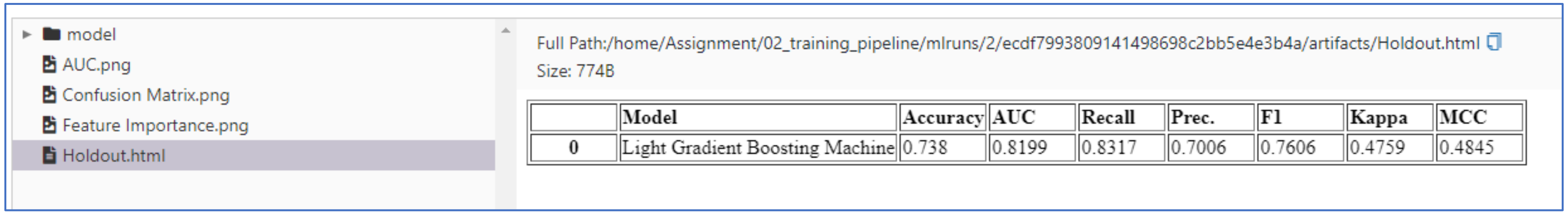
Predict on a Spark DataFrame:

```
import mlflow
logged_model = 'runs:/ecdf7993809141498698c2bb5e4e3b4a/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.
columns = list(df.columns)
df.withColumn('predictions', loaded_model(*columns)).collect()
```

Best Model Artifacts

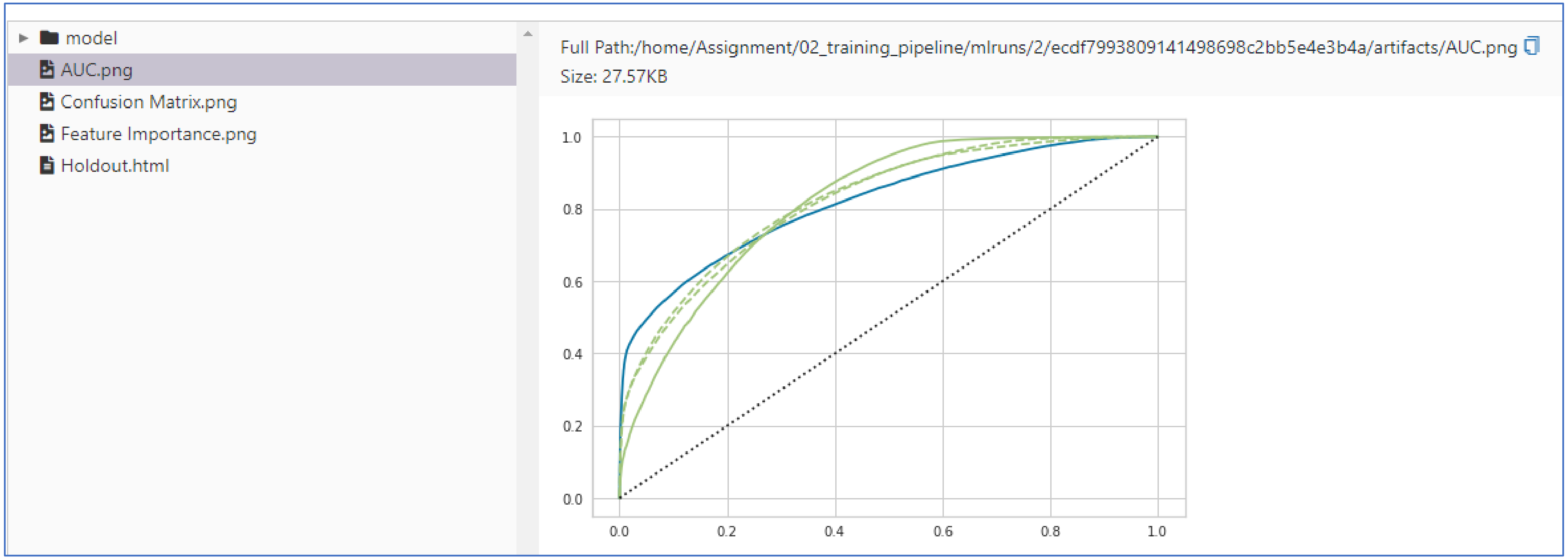


Full Path:/home/Assignment/02_training_pipeline/mlruns/2/ecdf7993809141498698c2bb5e4e3b4a/artifacts/Holdout.html

Size: 774B

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Light Gradient Boosting Machine	0.738	0.8199	0.8317	0.7006	0.7606	0.4759	0.4845

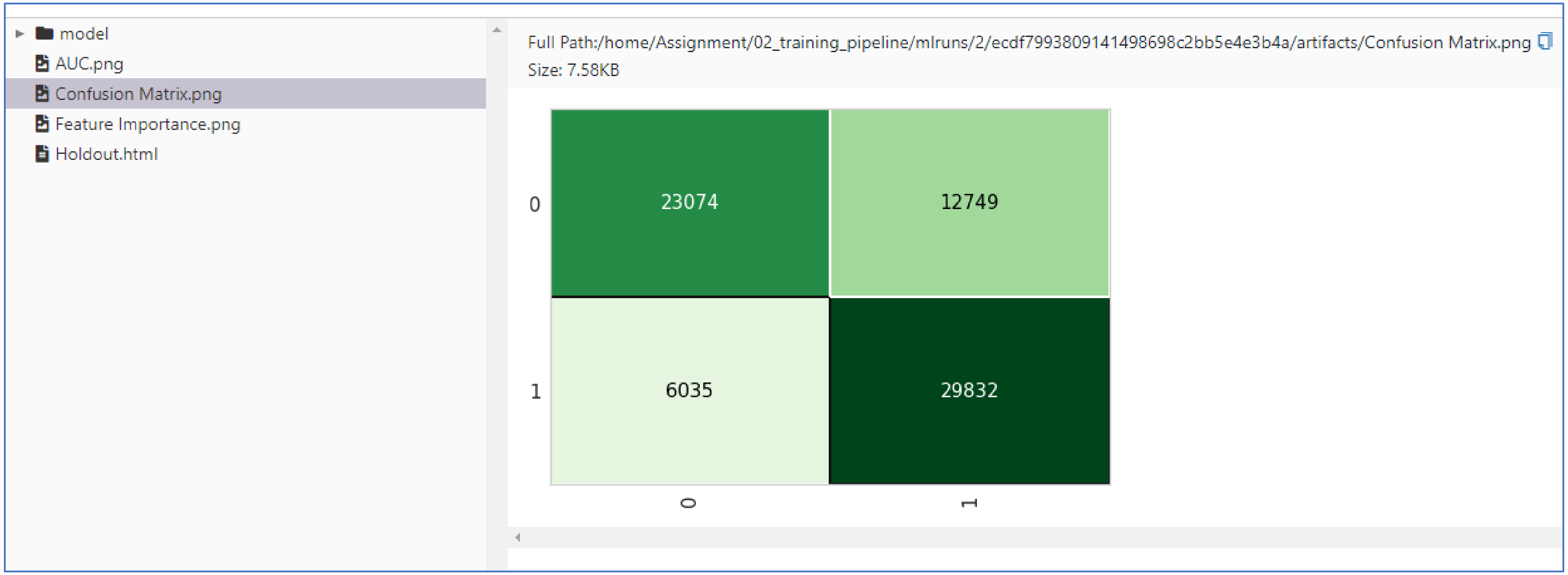
Best Model Artifacts



- model
- AUC.png
- Confusion Matrix.png
- Feature Importance.png
- Holdout.html

Full Path:/home/Assignment/02_training_pipeline/mlruns/2/ecdf7993809141498698c2bb5e4e3b4a/artifacts/AUC.png

Size: 27.57KB



- model
- AUC.png
- Confusion Matrix.png
- Feature Importance.png
- Holdout.html

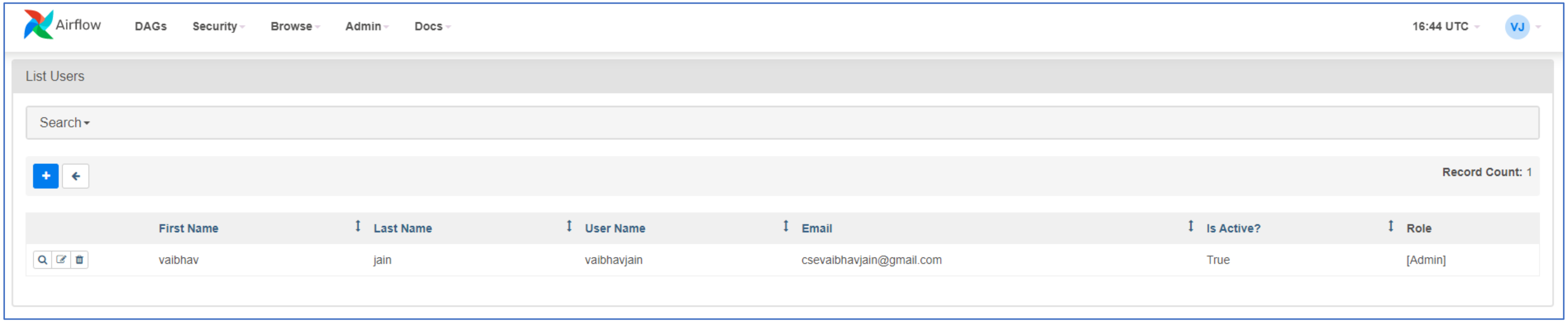
Full Path:/home/Assignment/02_training_pipeline/mlruns/2/ecdf7993809141498698c2bb5e4e3b4a/artifacts/Confusion Matrix.png

Size: 7.58KB

	0	1
0	23074	12749
1	6035	29832

Data Pipeline - Airflow UI

My User – Vaibhav Jain



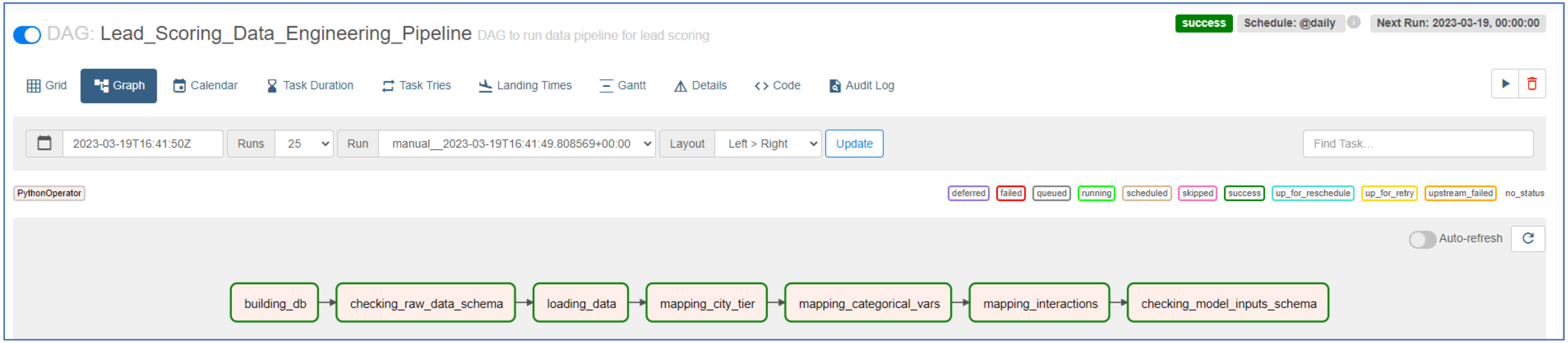
List Users

Search

Record Count: 1

	First Name	Last Name	User Name	Email	Is Active?	Role
	vaibhav	jain	vaibhavjain	csevaibhavjain@gmail.com	True	[Admin]

Pipeline Graph view

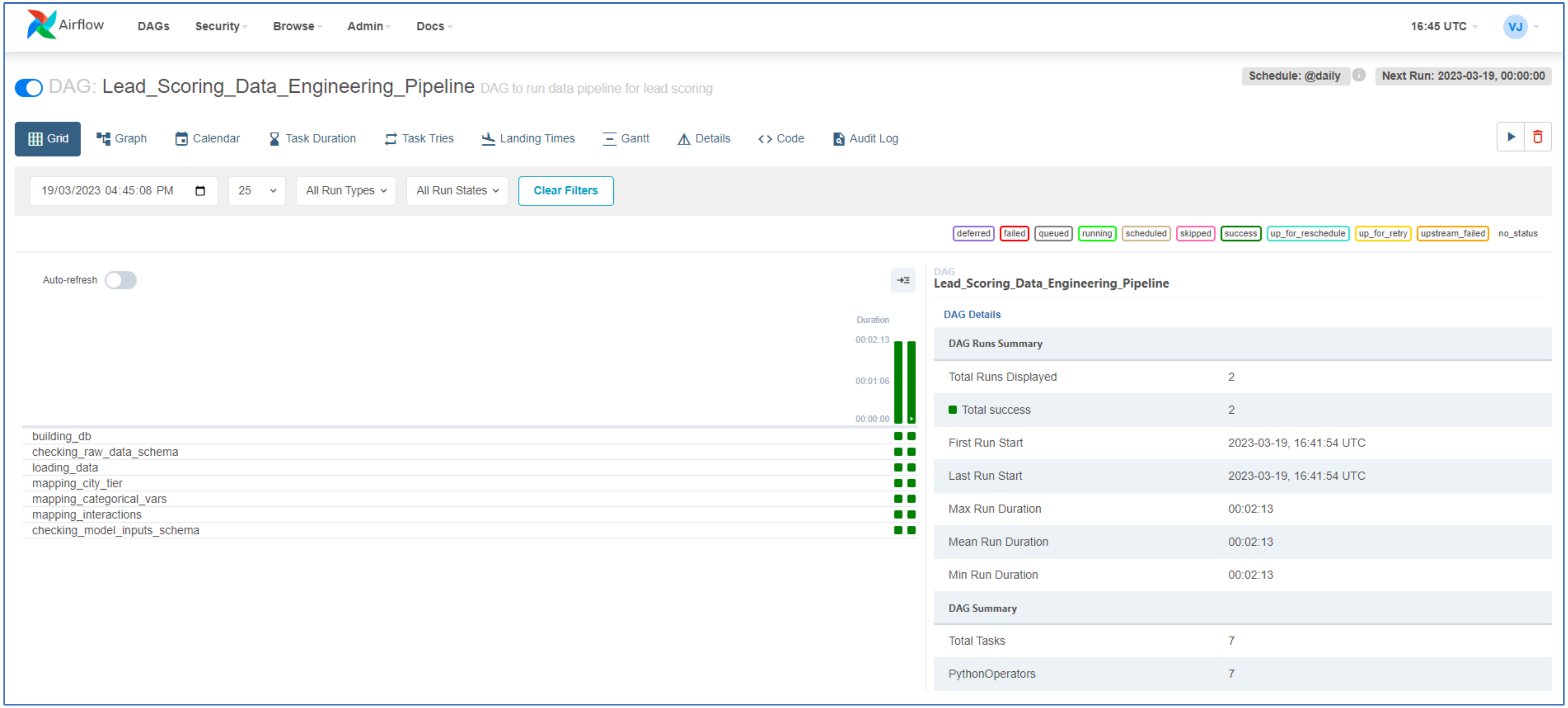


DAG: Lead_Scoring_Data_Engineering_Pipeline DAG to run data pipeline for lead scoring

success Schedule: @daily Next Run: 2023-03-19, 00:00:00

2023-03-19T16:41:50Z Runs 25 Run manual__2023-03-19T16:41:49.808569+00:00 Layout Left > Right

Pipeline Grid view



DAG: Lead_Scoring_Data_Engineering_Pipeline DAG to run data pipeline for lead scoring

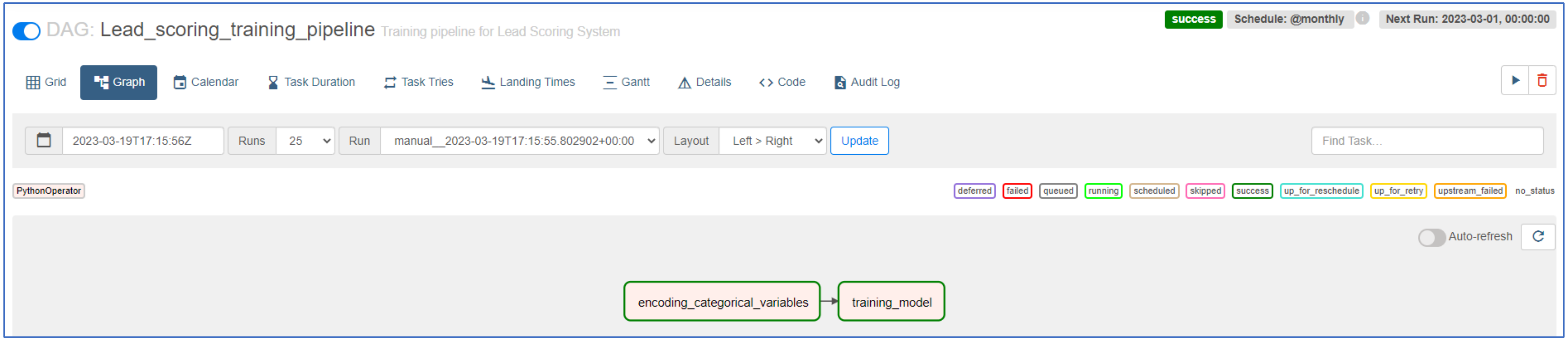
19/03/2023 04:45:08 PM

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 Runs Summary	
Total Runs Displayed	2
🟩 Total success	2
First Run Start	2023-03-19, 16:41:54 UTC
Last Run Start	2023-03-19, 16:41:54 UTC
Max Run Duration	00:02:13
Mean Run Duration	00:02:13
Min Run Duration	00:02:13
DAG Summary	
Total Tasks	7
PythonOperators	7

Training Pipeline – Airflow UI

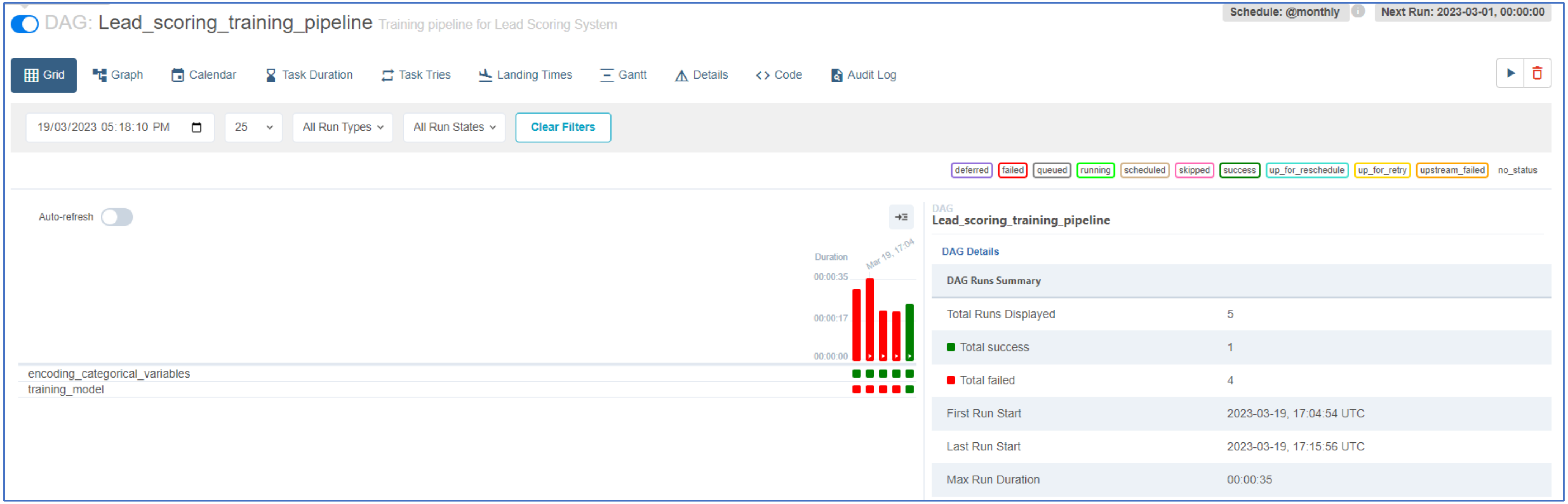


DAG: Lead_scoring_training_pipeline Training pipeline for Lead Scoring System

success Schedule: @monthly Next Run: 2023-03-01, 00:00:00

2023-03-19T17:15:56Z Runs 25 Run manual__2023-03-19T17:15:55.802902+00:00 Layout Left > Right Update

Pipeline Grid view



DAG: Lead_scoring_training_pipeline Training pipeline for Lead Scoring System

Schedule: @monthly Next Run: 2023-03-01, 00:00:00

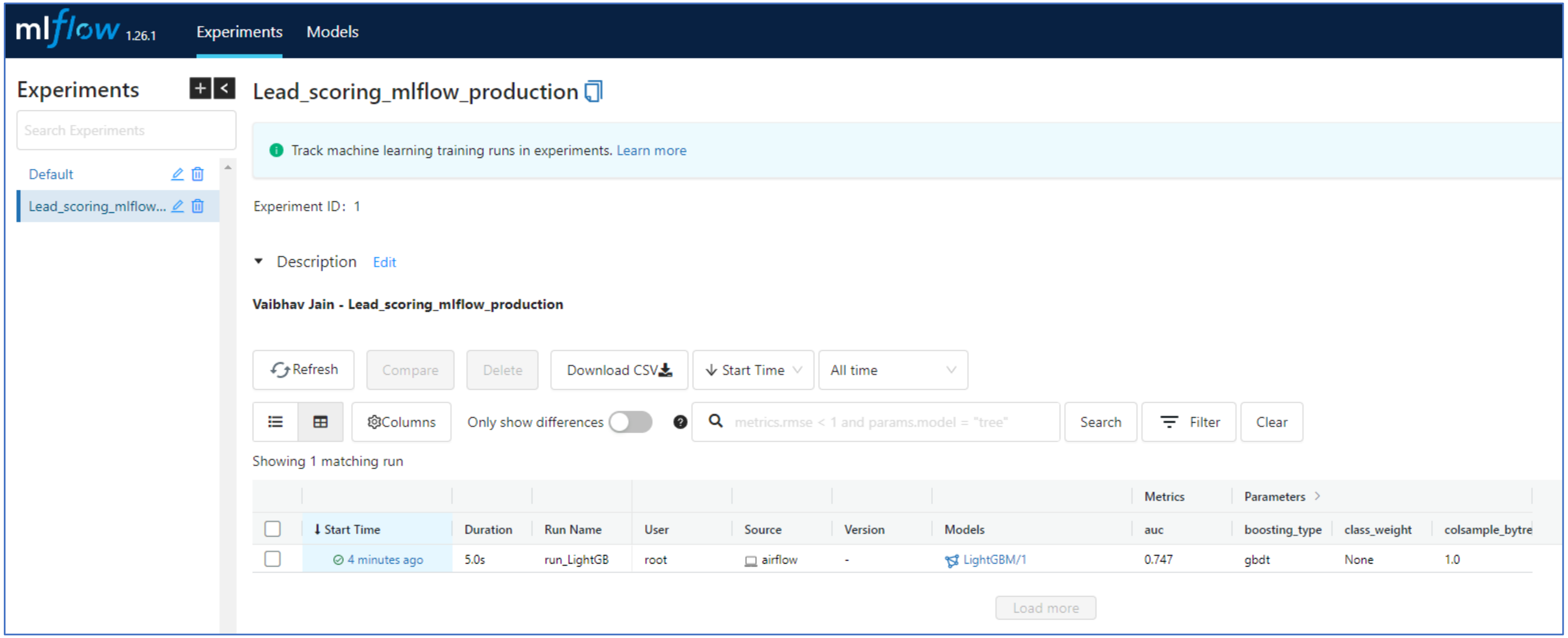
19/03/2023 05:18:10 PM 25 All Run Types All Run States Clear Filters

DAG
Lead_scoring_training_pipeline

DAG Details

DAG Runs Summary	
Total Runs Displayed	5
🟩 Total success	1
🟥 Total failed	4
First Run Start	2023-03-19, 17:04:54 UTC
Last Run Start	2023-03-19, 17:15:56 UTC
Max Run Duration	00:00:35

MLFlow – Production DB



Lead_scoring_mlflow_production

Track machine learning training runs in experiments. Learn more

Experiment ID: 1

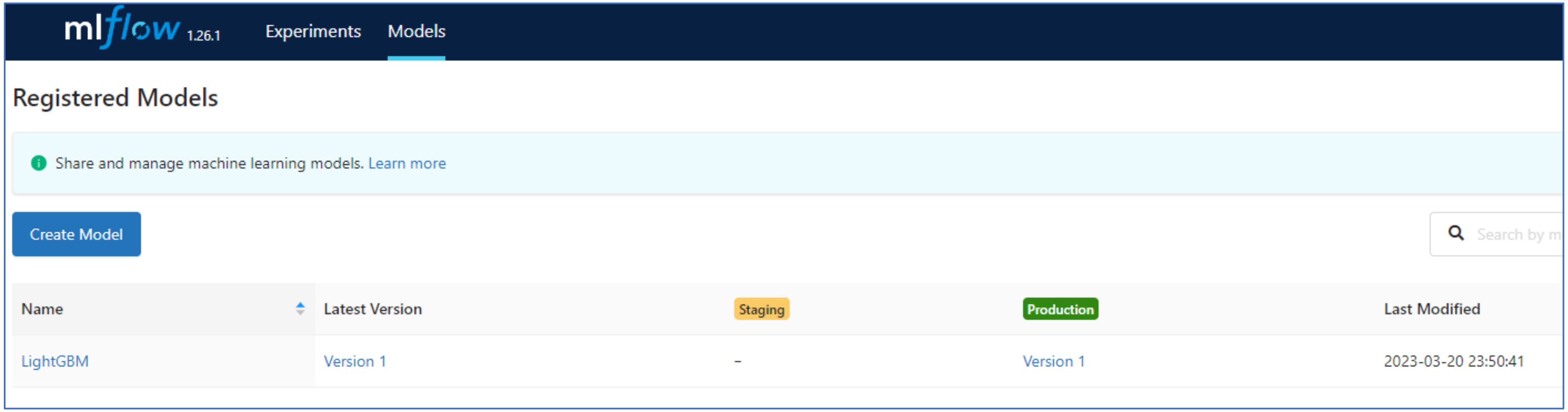
Description Edit

Vaibhav Jain - Lead_scoring_mlflow_production

Showing 1 matching run

								Metrics	Parameters		
□	Start Time	Duration	Run Name	User	Source	Version	Models	auc	boosting_type	class_weight	colsample_bytre
□	4 minutes ago	5.0s	run_LightGB	root	airflow	-	LightGBM/1	0.747	gbdt	None	1.0

Model staged as Production manually via MLFlow UI

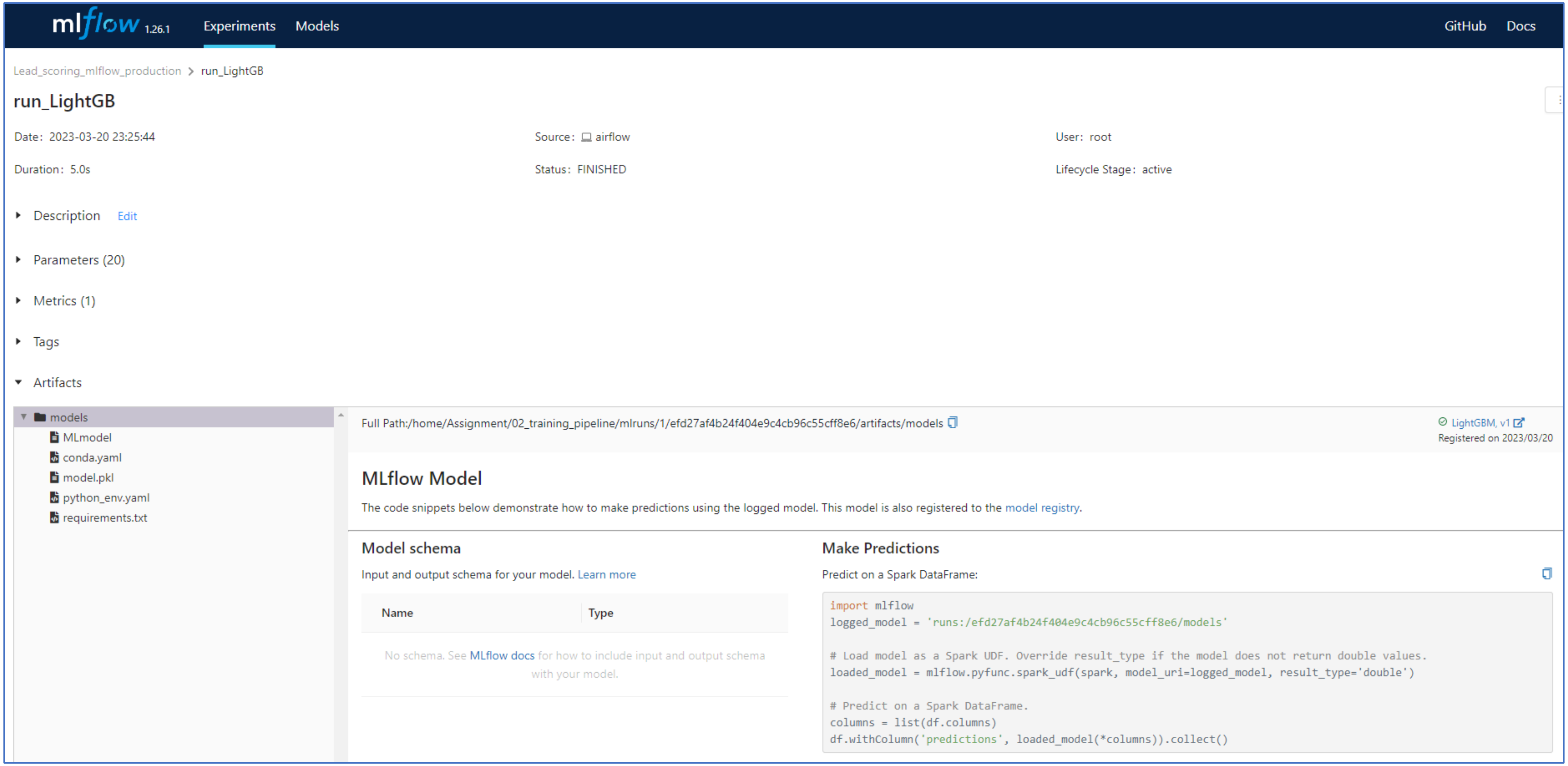


Registered Models

Share and manage machine learning models. Learn more

Name	Latest Version	Staging	Production	Last Modified
LightGBM	Version 1	-	Version 1	2023-03-20 23:50:41

MLFlow - Production DB - Model Artifacts



Lead_scoring_mlflow_production > run_LightGB

run_LightGB

Date: 2023-03-20 23:25:44

Source: airflow

User: root

Duration: 5.0s

Status: FINISHED

Lifecycle Stage: active

Full Path:/home/Assignment/02_training_pipeline/mlruns/1/efd27af4b24f404e9c4cb96c55cff8e6/artifacts/models

LightGBM, v1
Registered on 2023/03/20

MLflow Model

The code snippets below demonstrate how to make predictions using the logged model. This model is also registered to the model registry.

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

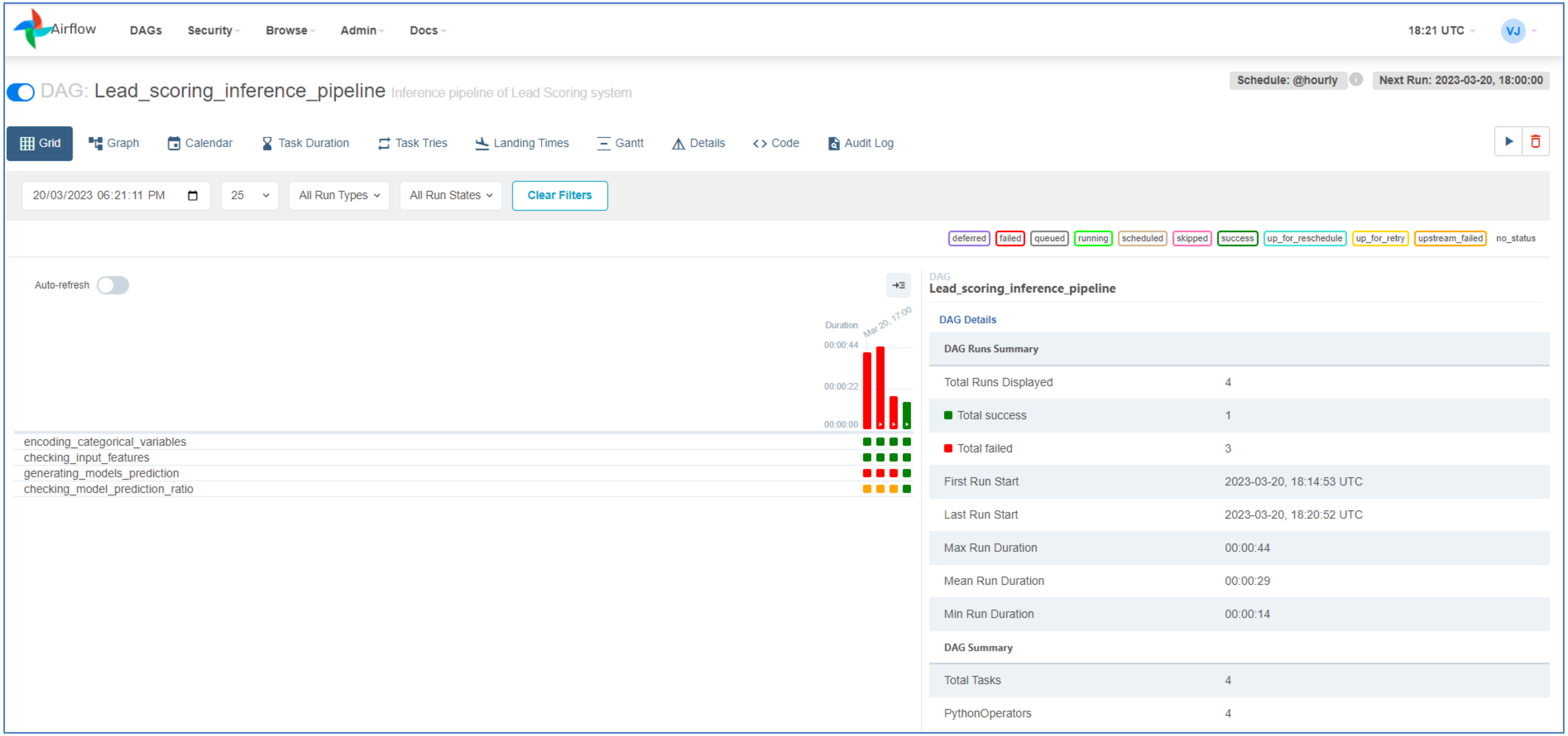
Predict on a Spark DataFrame:

```
import mlflow
logged_model = 'runs:/efd27af4b24f404e9c4cb96c55cff8e6/models'

# 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.
columns = list(df.columns)
df.withColumn('predictions', loaded_model(*columns)).collect()
```

Inference Pipeline – Airflow UI



DAG: Lead_scoring_inference_pipeline Inference pipeline of Lead Scoring system

Schedule: @hourly

Next Run: 2023-03-20, 18:00:00

20/03/2023 06:21:11 PM

25

All Run Types

All Run States

encoding_categorical_variables	🟩 🟩 🟩 🟩
checking_input_features	🟩 🟩 🟩 🟩
generating_models_prediction	🟥 🟥 🟥 🟩
checking_model_prediction_ratio	🟧 🟧 🟧 🟩

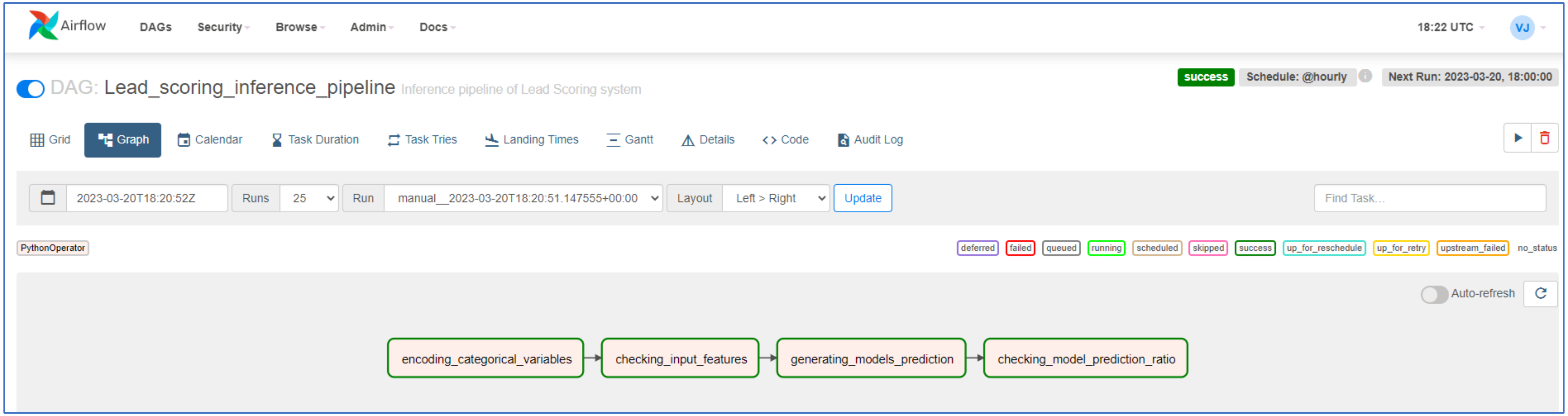
DAG

Lead_scoring_inference_pipeline

DAG Details

DAG Runs Summary	
Total Runs Displayed	4
🟩 Total success	1
🟥 Total failed	3
First Run Start	2023-03-20, 18:14:53 UTC
Last Run Start	2023-03-20, 18:20:52 UTC
Max Run Duration	00:00:44
Mean Run Duration	00:00:29
Min Run Duration	00:00:14

DAG Summary	
Total Tasks	4
PythonOperators	4



DAG: Lead_scoring_inference_pipeline Inference pipeline of Lead Scoring system

success Schedule: @hourly Next Run: 2023-03-20, 18:00:00

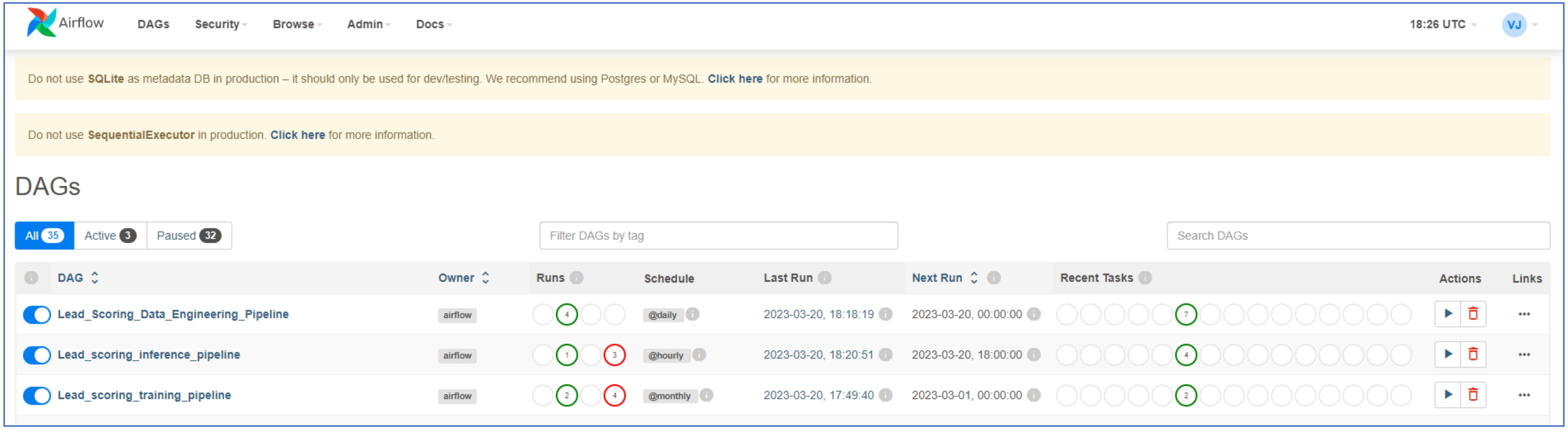
2023-03-20T18:20:52Z Runs 25 Run manual__2023-03-20T18:20:51.147555+00:00 Layout Left > Right

PythonOperator

deferred failed queued running scheduled skipped success up_for_reschedule up_for_retry upstream_failed no_status

encoding_categorical_variables → checking_input_features → generating_models_prediction → checking_model_prediction_ratio

Airflow – 3 dags ready



Do not use **SQLite** as metadata DB in production – it should only be used for dev/testing. We recommend using Postgres or MySQL. **Click here** for more information.

Do not use **SequentialExecutor** in production. **Click here** for more information.

DAGs

All 35 Active 3 Paused 32

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
Lead_Scoring_Data_Engineering_Pipeline	airflow	4	@daily	2023-03-20, 18:18:19	2023-03-20, 00:00:00	7	▶ 🗑	•••
Lead_scoring_inference_pipeline	airflow	1 3	@hourly	2023-03-20, 18:20:51	2023-03-20, 18:00:00	4	▶ 🗑	•••
Lead_scoring_training_pipeline	airflow	2 4	@monthly	2023-03-20, 17:49:40	2023-03-01, 00:00:00	2	▶ 🗑	•••