

Data Pipeline:



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 

DAG Lead_Scoring_Data_Engineering_Pipeline

DAG Details

DAG Summary

Total Tasks

7

PythonOperators

7

```
building_db
checking_raw_data_schema
loading_data
mapping_city_tier
mapping_categorical_vars
mapping_interactions
checking_model_inputs_schema
```



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

The Gantt chart displays tasks along the vertical axis and time along the horizontal axis. Tasks include: building_db, checking_raw_data_schema, loading_data, mapping_city_tier, mapping_categorical_vars, mapping_interactions, and checking_model_inputs_schema. A red bar indicates a failed task, while green bars indicate successful tasks. The chart shows runs starting at 00:00:00 and ending at 00:04:56 on Aug 21, 00:00.

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



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



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

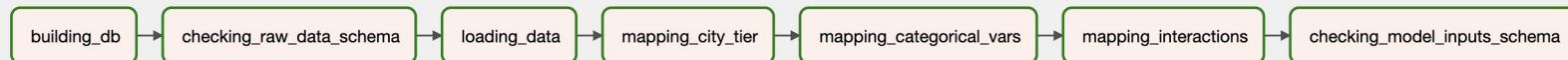
up_for_reschedule

up_for_retry

upstream_failed

no_status

Auto-refresh



Experiment on the cleaned data

The screenshot shows the mlflow UI interface. At the top, there is a dark header bar with the mlflow logo (1.26.1), navigation links for 'Experiments' and 'Models', and links for 'GitHub' and 'Docs'. Below the header is a sub-header 'Experiments' with a '+' icon, a back arrow, and the text 'Default' with a copy icon. To the right is a 'Share' button.

The main content area is titled 'Default' and contains a search bar labeled 'Search Experiments'. Below the search bar, there is a list of experiments: 'Default' (selected) and 'Lead_scoring_mlflow_...'. Each experiment entry has a copy icon and a delete icon.

Below the experiment list is a section titled 'Experiment ID: 0' with a 'Description' and an 'Edit' link. Underneath this are several buttons: 'Refresh', 'Compare', 'Delete', 'Download CSV', 'Start Time' dropdown set to 'All time', and a 'Columns' filter. There is also a toggle switch for 'Only show differences' and a search bar with a query 'metrics.rmse < 1 and params.model = "tree"'. To the right of the search bar are 'Search', 'Filter', and 'Clear' buttons.

The main table area displays a message 'Showing 0 matching runs' and a table header with columns: 'Start Time' (sorted by start time), 'Duration', 'Run Name', 'User', 'Source', 'Version', and 'Models'. A note at the bottom states 'No runs yet. Learn more about how to create ML model training runs in this experiment.'

Experiments

+ < Lead_scoring_mlflow_production [Share](#)

Search Experiments

X

Default



Lead_scoring_mlflow_...



Experiment ID: 1

▶ Description [Edit](#)

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

	↓ Start Time	Duration	Run Name	User	Source	Version	Models	AUC	Accuracy	F1	C	CPU Jobs	Cat
<input type="checkbox"/>	19 minutes ago		Session Init...	root		-	-	-	-	-	-	-1	4
<input type="checkbox"/>	3 minutes ago		Light Gradi...	root		-		0.814	0.736	0.759	-	-	-
<input type="checkbox"/>	3 minutes ago		Light Gradi...	root		-		0.815	0.736	0.76	-	-	-
<input type="checkbox"/>	12 minutes ago		Light Gradi...	root		-		0.815	0.736	0.76	-	-	-
<input type="checkbox"/>	13 minutes ago		Naive Bayes	root		-		0.722	0.661	0.721	-	-	-
<input type="checkbox"/>	13 minutes ago		Linear Disc...	root		-		0.762	0.696	0.725	-	-	-
<input type="checkbox"/>	13 minutes ago		Ridge Clas...	root		-		0	0.696	0.725	-	-	-
<input type="checkbox"/>	13 minutes ago		Logistic Re...	root		-		0.776	0.707	0.737	1.0	-	-
<input type="checkbox"/>	13 minutes ago		Random Fo...	root		-		0.811	0.732	0.759	-	-	-
<input type="checkbox"/>	13 minutes ago		Extra Trees...	root		-		0.812	0.733	0.756	-	-	-
<input type="checkbox"/>	13 minutes ago		Decision Tr...	root		-		0.811	0.733	0.756	-	-	-
<input type="checkbox"/>	13 minutes ago		Extreme Gr...	root		-		0.814	0.736	0.759	-	-	-
<input type="checkbox"/>	13 minutes ago		Light Gradi...	root		-		0.815	0.736	0.76	-	-	-
<input type="checkbox"/>	16 minutes ago		Session Init...	root		-	-	-	-	-	-	-1	4
<input type="checkbox"/>	16 minutes ago		MLflow Model	root		-		0.811	0.732	0.756	-	-	-

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

▼ Tags (5)

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

Name Value

▼ Artifacts

LightGBM model

Full Path:/home/Assignment/02_training_pipeline/notebooks/mlruns/1/9884b4262c7c4c3ebe147a7540d03aa2/artifacts/LightGBM 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')
```

▶ Description [Edit](#)

▶ Parameters (21)

▶ Metrics (8)

▶ Tags (5)

▼ Artifacts

LightGBM model

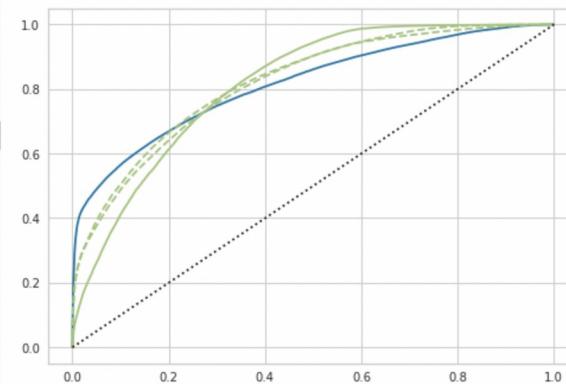
- MLmodel
- conda.yaml
- model.pkl
- python_env.yaml
- requirements.txt

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 [Download](#)

Size: 27.87KB



▶ Description [Edit](#)

▶ Parameters (21)

▶ Metrics (8)

▶ Tags (5)

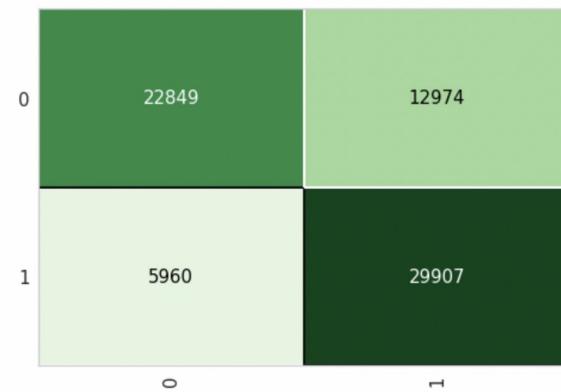
▼ Artifacts

▶ LightGBM model

- MLmodel
- conda.yaml
- model.pkl
- python_env.yaml
- requirements.txt

Full Path:/home/Assignment/02_training_pipeline/notebooks/mlruns/1/9884b4262c7c4c3ebe147a7540d03aa2/artifacts/Confusion Matrix.p...

Size: 7.9KB



▶ model

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

Confusion Matrix.png

► Description [Edit](#)

► Parameters (21)

► Metrics (8)

► Tags (5)

▼ Artifacts

LightGBM model

- MLmodel
- conda.yaml
- model.pkl
- python_env.yaml
- requirements.txt

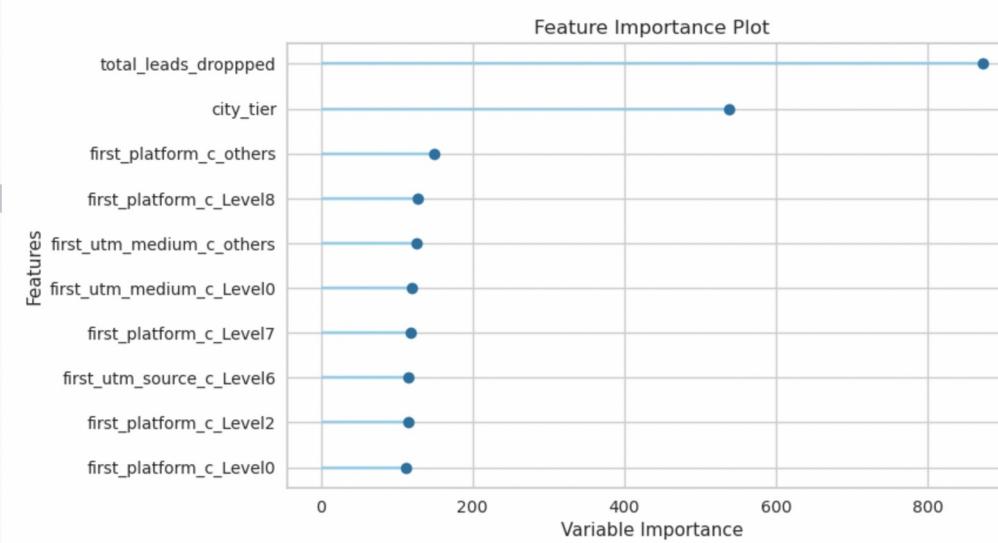
model

- AUC.png
- Confusion Matrix.png

[Feature Importance.png](#)

Holdout.html

Full Path:/home/Assignment/02_training_pipeline/notebooks/mlruns/1/9884b4262c7c4c3ebe147a7540d03aa2/artifacts/Feature Importanc... [Download](#)
Size: 41.03KB



► Description [Edit](#)

► Parameters (21)

► Metrics (8)

► Tags (5)

▼ Artifacts

▼ LightGBM model

MLmodel
 conda.yaml
 model.pkl
 python_env.yaml
 requirements.txt

Full Path:/home/Assignment/02_training_pipeline/notebooks/mlruns/1/9884b4262c7c4c3ebe147a7540d03aa2/artifacts/Holdout.html [🔗](#)

Size: 774B



	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

► model

AUC.png
 Confusion Matrix.png
 Feature Importance.png

Holdout.html

Registering and Promoting model to Production

The screenshot shows the mlflow UI interface for managing machine learning models. At the top, there is a navigation bar with the mlflow logo (1.26.1), Experiments, Models (selected), GitHub, and Docs.

The main area is titled "Registered Models". It includes a help message: "Share and manage machine learning models. [Learn more](#)".

Below the message are buttons for "Create Model" and search/filter options: "Search by model name" (with a magnifying glass icon), "Search", "Filter", and "Clear".

A table lists registered models:

Name	Latest Version	Staging	Production	Last Modified	Tags
LightGBM	Version 5	-	-	2024-08-26 09:11:41	-

Pagination controls at the bottom right include icons for previous/next pages and a dropdown for "10 / page".

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

mlflow 1.26.1 Experiments Models GitHub Docs

Registered Models > LightGBM > Version 5

Version 5

Registered At: 2024-08-26 09:11:41 Stage: **Staging** ▾ Last Modified: 2024-08-26 09:22:32

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 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: **Production** ▾ Last Modified: 2024-08-26 09:22:49

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.

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](#)

<input type="checkbox"/>	Version	Registered at	Created by	Stage	Description
<input type="checkbox"/>	Version 5	2024-08-26 09:11:41		Production	
<input type="checkbox"/>	Version 4	2024-08-26 09:07:47		None	
<input type="checkbox"/>	Version 3	2024-08-26 09:07:04		None	
<input type="checkbox"/>	Version 2	2024-08-26 09:01:33		None	
<input type="checkbox"/>	Version 1	2024-08-26 08:58:53		None	

Training Pipeline



Airflow

DAGs

Security

Browse

Admin

Docs

11:41 UTC

UU

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 34 Active 1 Paused 33

Filter DAGs by tag

Search DAGs

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
Lead_Scoring_Data_Engineering_Pipeline	airflow	2 3	@daily	2024-08-22, 07:46:51	2024-08-22, 00:00:00	7		
Lead_scoring_training_pipeline	airflow	4	@monthly		2024-07-01, 00:00:00	7		
example_bash_operator example example2	airflow	4	0 0 * * *		2024-08-21, 00:00:00	7		
example_branch_datetime_operator example	airflow	4	@daily		2024-08-21, 00:00:00	7		
example_branch_datetime_operator_2 example	airflow	4	@daily		2024-08-21, 00:00:00	7		
example_branch_dop_operator_v3 example	airflow	4	*/* * * *		2024-08-22, 11:38:00	7		
example_branch_labels	airflow	4	@daily		2024-08-21, 00:00:00	7		
example_branch_operator example example2	airflow	4	@daily		2024-08-21, 00:00:00	7		
example_branch_python_operator_decorator example example2	airflow	4	@daily		2024-08-21, 00:00:00	7		
example_complex example example2 example3	airflow	4	None			7		



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

DAG
Lead_scoring_training_pipeline

DAG Details

DAG Summary

Total Tasks	2
PythonOperators	2

encoding_categorical_variables
training_model



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

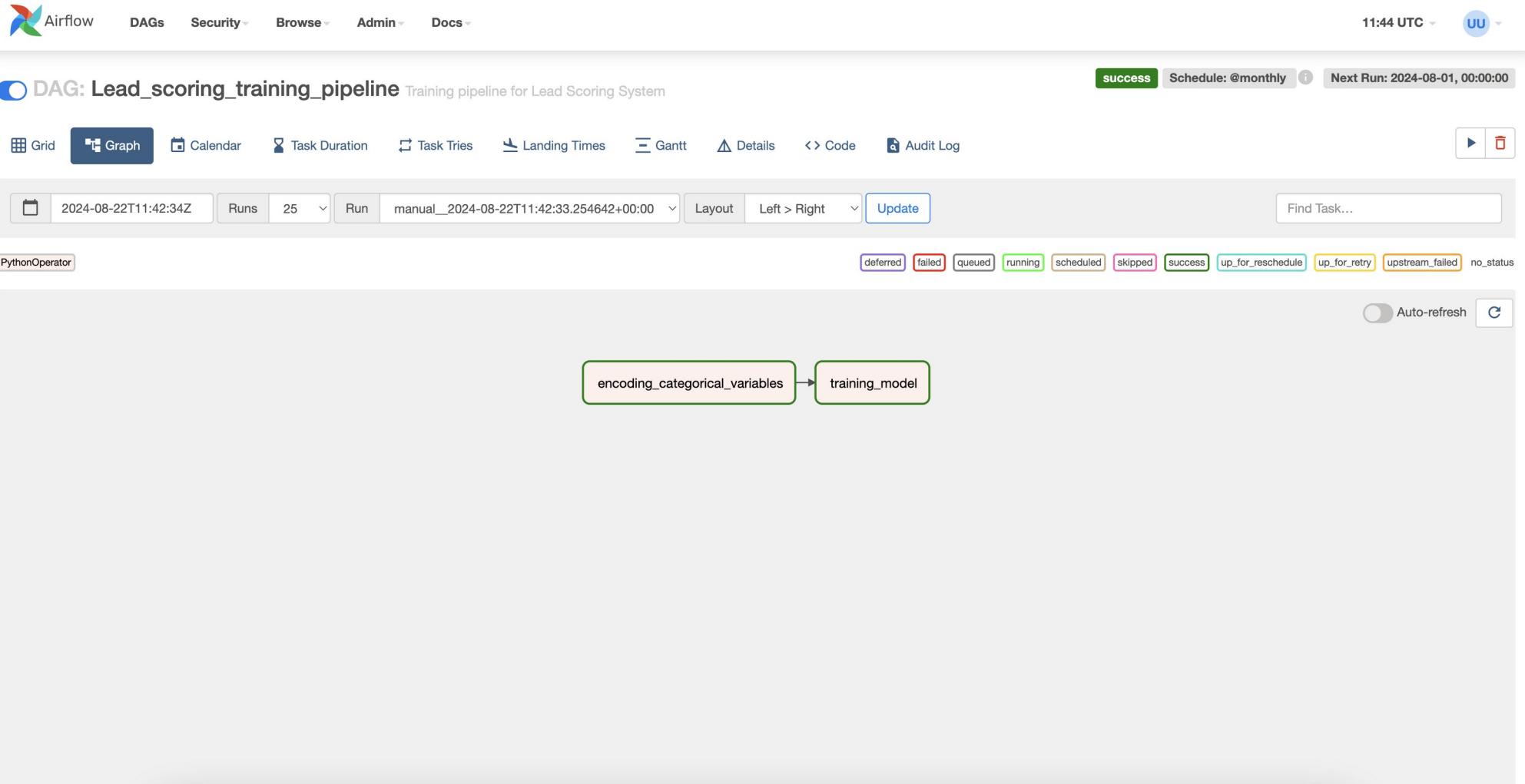
DAG
Lead_scoring_training_pipeline

DAG Details

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



Inference Pipeline

Airflow DAGs Security Browse Admin Docs 08:31 UTC UU

DAG: Lead_scoring_inference_pipeline Inference pipeline of Lead Scoring system

Schedule: @hourly Next Run: 2024-08-26, 08:00:00

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

26/08/2024, 08:31:08 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
00:01:25
00:00:42
00:00:00

encoding_categorical_variables
generating_models_prediction
checking_model_prediction_ratio
checking_input_features

DAG
Lead_scoring_inference_pipeline

DAG Details

Total Runs Displayed	2
Total success	2
First Run Start	2024-08-26, 08:28:55 UTC
Last Run Start	2024-08-26, 08:28:59 UTC
Max Run Duration	00:01:25
Mean Run Duration	00:01:25
Min Run Duration	00:01:25

DAG Summary

Total Tasks	4
PythonOperators	4



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

