

## Airflow-UI showing all the 3 pipelines

Do not use **SQLite** as metadata DB in production – it should only be used for devtesting. We recommend using Postgres or MySQL. [Click here](#) for more information.

Do not use **SequentialExecutor** in production. [Click here](#) for more information.

### DAGs

Filter DAGs by tag  Search DAGs

| DAG  | Owner   | Runs                          | Schedule  | Last Run             | Next Run             | Recent Tasks                  |
|--|---------|-------------------------------|-----------|----------------------|----------------------|-------------------------------|
| <input checked="" type="checkbox"/> Lead_Scoring_Data_Engineering_Pipeline | airflow | <span>5</span> <span>4</span> | @daily    | 2024-06-08, 10:57:47 | 2024-06-09, 00:00:00 | <span>5</span> <span>4</span> |
| <input checked="" type="checkbox"/> Lead_scoring_inference_pipeline        | airflow | <span>1</span> <span>3</span> | @hourly   | 2024-06-09, 13:43:34 | 2024-06-09, 13:00:00 | <span>4</span> <span>3</span> |
| <input checked="" type="checkbox"/> Lead_scoring_training_pipeline         | airflow | <span>5</span> <span>1</span> | @monthly  | 2024-06-09, 10:43:46 | 2024-07-01, 00:00:00 | <span>2</span> <span>1</span> |
| <input type="checkbox"/> example_branch_operator                           | airflow | <span>0</span> <span>0</span> | 0 8 * * * |                      | 2024-06-08, 00:00:00 | <span>0</span> <span>0</span> |
| <input type="checkbox"/> example_branch_datetime_operator                  | airflow | <span>0</span> <span>0</span> | @daily    |                      | 2024-06-08, 00:00:00 | <span>0</span> <span>0</span> |
| <input type="checkbox"/> example_branch_datetime_operator_2                | airflow | <span>0</span> <span>0</span> | @daily    |                      | 2024-06-08, 00:00:00 | <span>0</span> <span>0</span> |
| <input type="checkbox"/> example_branch_dop_operator_v3                    | airflow | <span>0</span> <span>0</span> |           |                      |                      | <span>0</span> <span>0</span> |

## Model Experimentation – MLFlow

### i. Baseline model

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](#) [Filter](#) [Clear](#)

Showing 35 matching runs

|                          | Start Time | Duration | Run Name         | User | Source    | Version | Models  | Metrics           | Parameters | Tags         |
|--------------------------|------------|----------|------------------|------|-----------|---------|---------|-------------------|------------|--------------|
| <input type="checkbox"/> | 1 day ago  | 49.5min  | Session Init...  | root | bykeme... | -       | -       | -                 | -          | -            |
| <input type="checkbox"/> | 7 days ago |          | Session Init...  | root | bykeme... | -       | -       | -                 | -          | -            |
| <input type="checkbox"/> | 7 days ago |          | Extreme Gra...   | root | bykeme... | -       | -       | -                 | -          | -            |
| <input type="checkbox"/> | 7 days ago |          | Naive Bayes      | root | bykeme... | -       | sklearn | 0.734 0.663 0.727 | -          | compare_m... |
| <input type="checkbox"/> | 7 days ago |          | Linear Disc...   | root | bykeme... | -       | sklearn | 0.773 0.701 0.728 | -          | compare_m... |
| <input type="checkbox"/> | 7 days ago |          | Ridge Class...   | root | bykeme... | -       | sklearn | 0 0.701 0.728     | -          | compare_m... |
| <input type="checkbox"/> | 7 days ago |          | Logistic Reg...  | root | bykeme... | -       | sklearn | 0.784 0.71 0.74   | 1.0        | compare_m... |
| <input type="checkbox"/> | 7 days ago |          | Random For...    | root | bykeme... | -       | sklearn | 0.817 0.735 0.761 | -          | compare_m... |
| <input type="checkbox"/> | 7 days ago |          | Decision Tre...  | root | bykeme... | -       | sklearn | 0.817 0.736 0.758 | -          | compare_m... |
| <input type="checkbox"/> | 7 days ago |          | Extra Trees C... | root | bykeme... | -       | sklearn | 0.816 0.737 0.758 | -          | compare_m... |
| <input type="checkbox"/> | 7 days ago |          | Light Gradie...  | root | bykeme... | -       | sklearn | 0.821 0.739 0.762 | -          | compare_m... |
| <input type="checkbox"/> | 7 days ago |          | Extreme Gra...   | root | bykeme... | -       | -       | -                 | -          | -            |
| <input type="checkbox"/> | 7 days ago |          | Session Init...  | root | bykeme... | -       | -       | -                 | -1         | 4            |
| <input type="checkbox"/> | 7 days ago |          | Session Init...  | root | bykeme... | -       | -       | -                 | -1         | 4            |
| <input type="checkbox"/> | 7 days ago |          | Session Init...  | root | bykeme... | -       | -       | -                 | -1         | 4            |

## ii. Experiment with features dropped

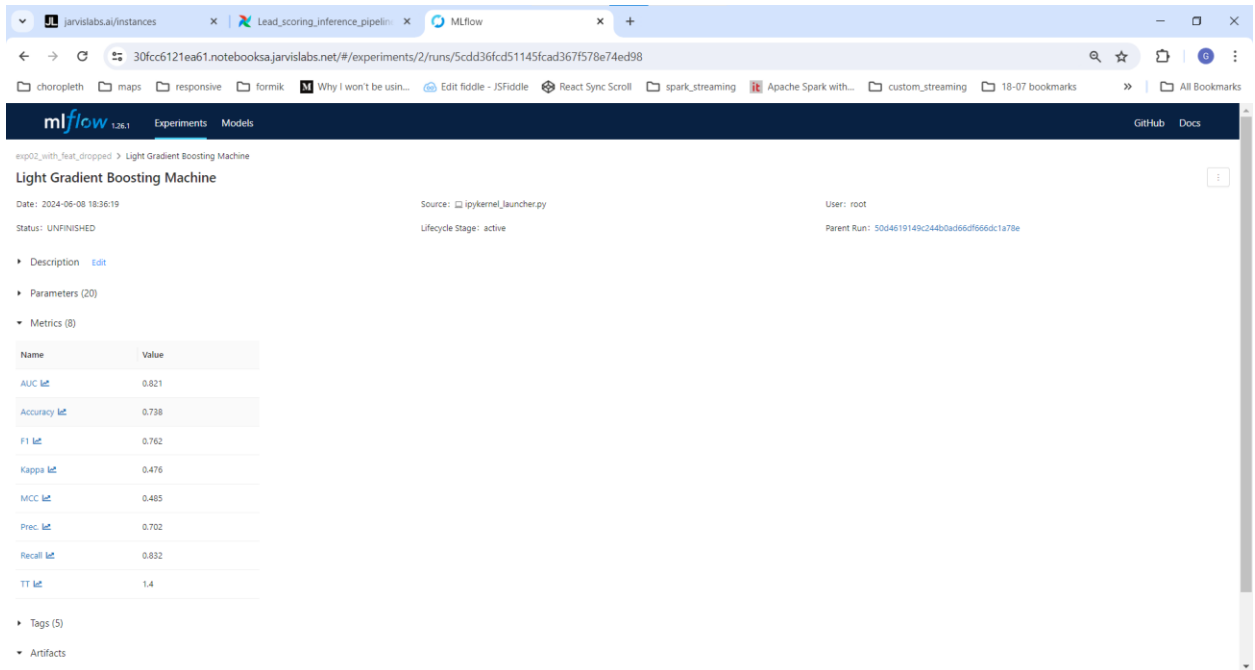
The screenshot shows the MLflow Experiments page for experiment 'exp02\_with\_feat\_dropped'. The interface includes a sidebar with experiment selection, a top navigation bar, and a main content area with a table of runs. The table columns include Start Time, Duration, Run Name, User, Source, Version, Models, Metrics (AUC, Accuracy, F1, C), CPU Jobs, Categorical Feat, Source, URI, and URI. The runs are sorted by Start Time, showing a sequence of model training runs with varying metrics.

| Start Time | Duration | Run Name         | User | Source       | Version | Models  | AUC   | Accuracy | F1    | C   | CPU Jobs | Categorical Feat | Source       | URI      | URI |
|------------|----------|------------------|------|--------------|---------|---------|-------|----------|-------|-----|----------|------------------|--------------|----------|-----|
| 1 day ago  |          | Session Init...  | root | ipykernel... | -       | -       | -     | -        | -     | -   | -1       | 5                | setup        | c89233c4 | ee4 |
| 1 day ago  |          | Naive Bayes      | root | ipykernel... | -       | sklearn | 0.734 | 0.672    | 0.725 | -   | -        | -                | company_m... | b0d36c7d | ee4 |
| 1 day ago  |          | Linear Discr...  | root | ipykernel... | -       | sklearn | 0.773 | 0.7      | 0.727 | -   | -        | -                | company_m... | 20143fab | ee4 |
| 1 day ago  |          | Ridge Classif... | root | ipykernel... | -       | sklearn | 0     | 0.7      | 0.727 | -   | -        | -                | company_m... | e5608a1b | ee4 |
| 1 day ago  |          | Logistic Reg...  | root | ipykernel... | -       | sklearn | 0.784 | 0.71     | 0.74  | 1.0 | -        | -                | company_m... | e590c26  | ee4 |
| 1 day ago  |          | Decision Tre...  | root | ipykernel... | -       | sklearn | 0.817 | 0.736    | 0.758 | -   | -        | -                | company_m... | afed6a11 | ee4 |
| 1 day ago  |          | Extra Trees C... | root | ipykernel... | -       | sklearn | 0.817 | 0.737    | 0.758 | -   | -        | -                | company_m... | 53646d40 | ee4 |
| 1 day ago  |          | Random For...    | root | ipykernel... | -       | sklearn | 0.818 | 0.737    | 0.759 | -   | -        | -                | company_m... | ae65d761 | ee4 |
| 1 day ago  |          | Light Gradie...  | root | ipykernel... | -       | sklearn | 0.821 | 0.738    | 0.762 | -   | -        | -                | company_m... | 30e093b6 | ee4 |
| 6 days ago |          | Session Init...  | root | ipykernel... | -       | -       | -     | -        | -     | -   | -1       | 5                | setup        | a9924e44 | da2 |
| 7 days ago |          | Session Init...  | root | ipykernel... | -       | -       | -     | -        | -     | -   | -1       | 5                | setup        | d7c82883 | 2b1 |
| 7 days ago |          | Session Init...  | root | ipykernel... | -       | -       | -     | -        | -     | -   | -1       | 5                | setup        | 070c4c4f | a4c |

## iii. Best model LGBM with parameters & metrics

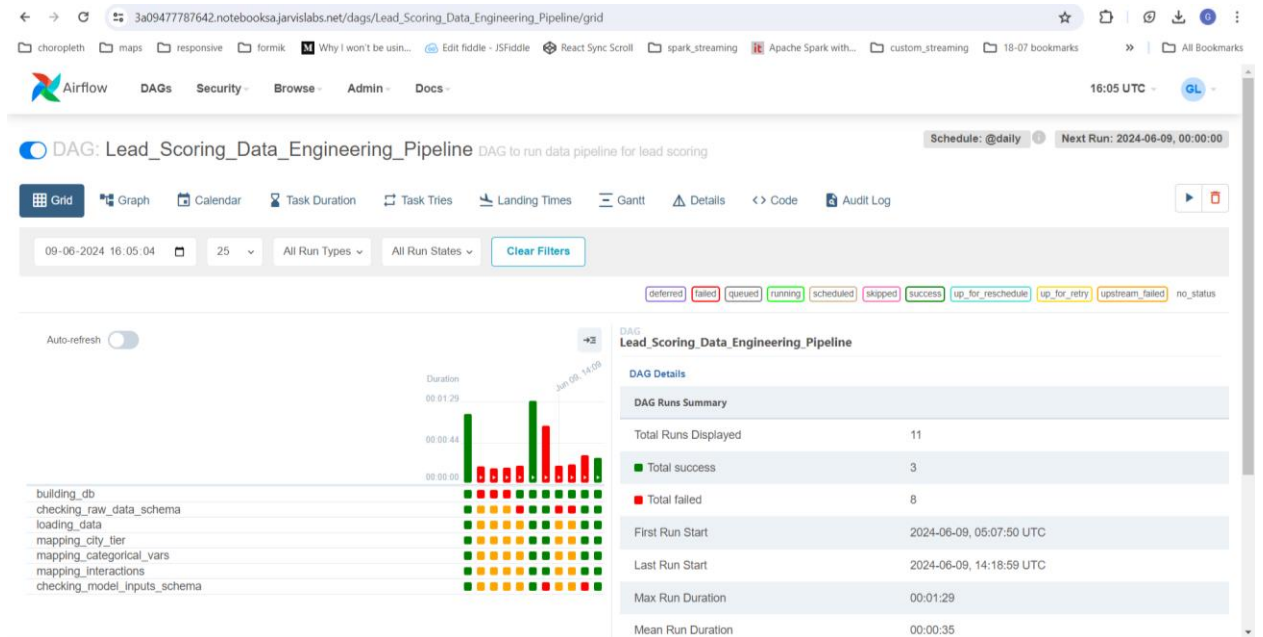
The screenshot shows the MLflow page for the best model, 'Light Gradient Boosting Machine'. The page displays the model's date, status, source, lifecycle stage, user, and parent run. Below this, the parameters and their values are listed in a table.

| Name              | Value |
|-------------------|-------|
| boosting_type     | gbdt  |
| class_weight      | None  |
| colsample_bytree  | 1.0   |
| 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        | 31    |



## Data Pipeline

### 1. Airflow UI Graph:



## 2. Airflow UI Grid:

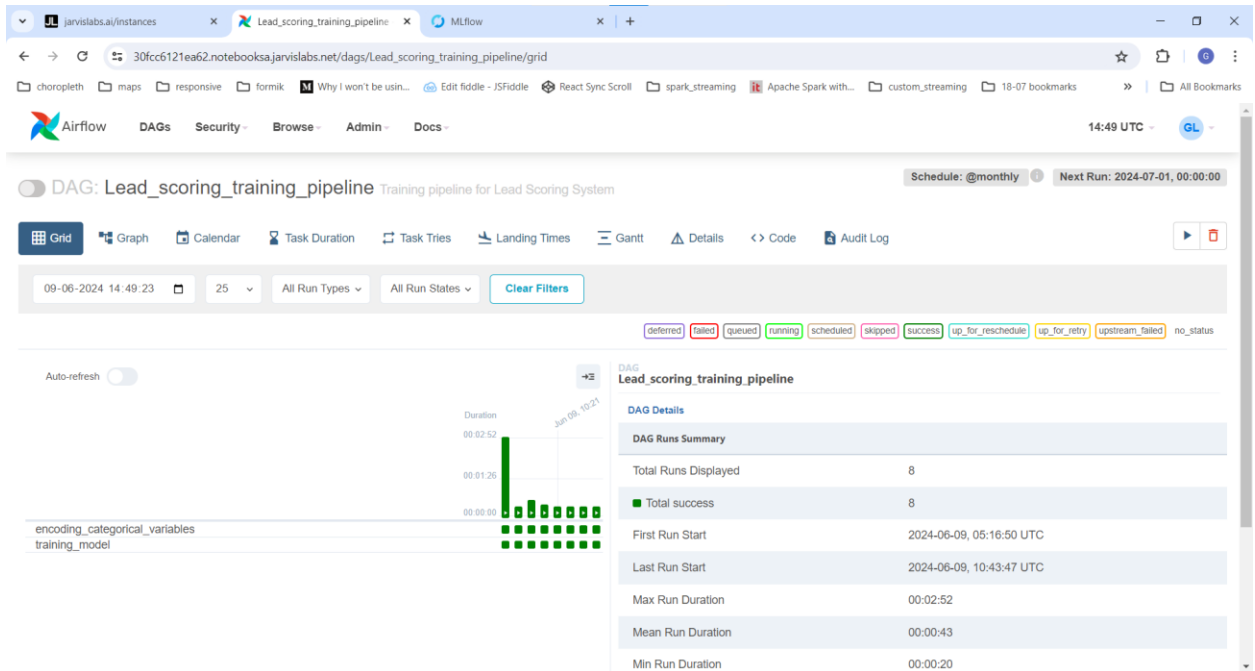
The screenshot shows the Airflow web interface for the DAG 'Lead\_Scoring\_Data\_Engineering\_Pipeline'. The top navigation bar includes 'Airflow', 'DAGs', 'Security', 'Browse', 'Admin', and 'Docs'. The DAG status is 'success' with a schedule of '@daily' and a next run of '2024-06-09, 00:00:00'. The 'Grid' tab is selected, showing a table of DAG runs. The first run is highlighted, showing a start time of '2024-06-09T14:18:58Z' and a status of 'success'. Below the table, a task graph is displayed, showing a sequence of tasks: 'building\_db', 'checking\_raw\_data\_schema', 'loading\_data', 'mapping\_city\_tier', 'mapping\_categorical\_vars', 'mapping\_interactions', and 'checking\_model\_inputs\_schema'. The tasks are connected by arrows, indicating a linear flow. The 'Auto-refresh' toggle is turned on.

## Training Pipeline

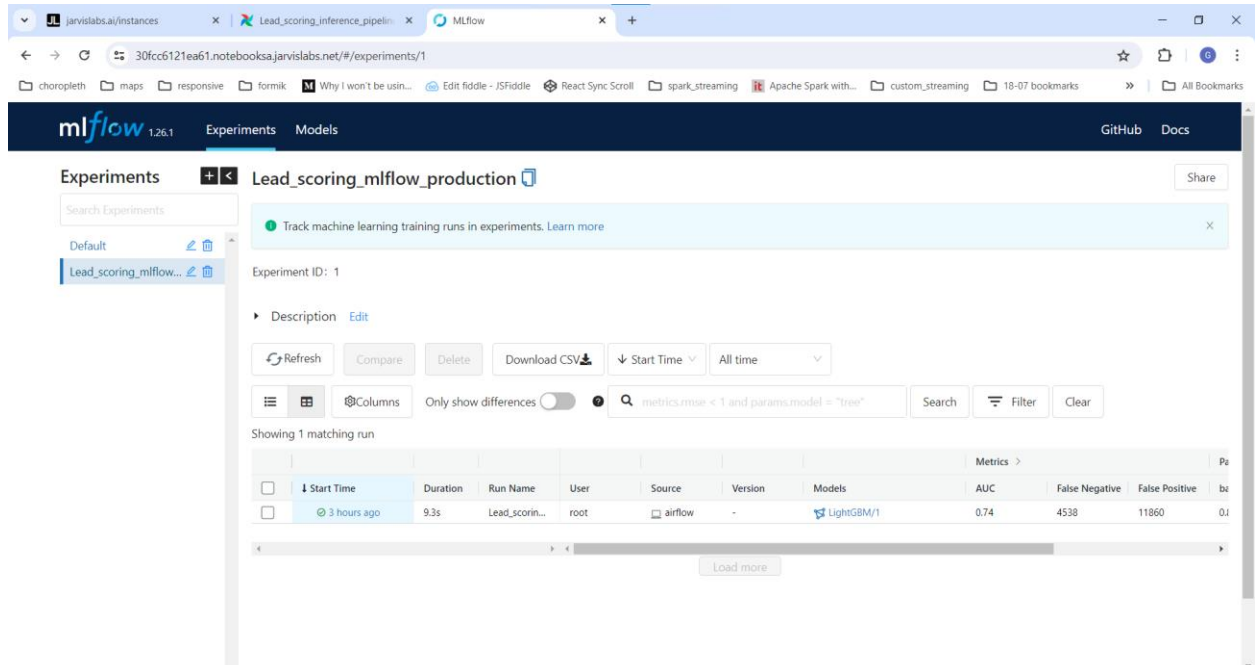
### 1. Airflow UI Graph:

The screenshot shows the Airflow web interface for the DAG 'Lead\_scoring\_training\_pipeline'. The top navigation bar includes 'Airflow', 'DAGs', 'Security', 'Browse', 'Admin', and 'Docs'. The DAG status is 'success' with a schedule of '@monthly' and a next run of '2024-07-01, 00:00:00'. The 'Graph' tab is selected, showing a task graph with two tasks: 'encoding\_categorical\_variables' and 'training\_model'. The tasks are connected by an arrow, indicating a linear flow. The 'Auto-refresh' toggle is turned on.

## 2. Airflow UI Grid:



## 3. MLFlow UI Experiments Page



#### 4. MLFlow UI with Artifacts visible

The screenshot shows the MLflow web interface for a registered model named 'LightGBM'. The page header includes the MLflow logo and navigation links for 'Experiments' and 'Models'. The 'Registered Models' breadcrumb is visible. The model name 'LightGBM' is prominently displayed. Below the name, the creation time '2024-06-09 16:14:06' and last modified time '2024-06-09 18:57:52' are shown. The 'Description' tab is selected, and the 'Tags' section is empty. The 'Versions' section shows a table with one version, 'Version 1', which is the active version and is in 'Production' status.

## 5. MLFlow UI Model Promoted

javaslabs.com

Lead\_scoring\_inference\_pipeline

Mflow

30fcc6121ea61.notebooks.javaslabs.net/#/experiments/1/runs/2b94292f9def480c8469832dd7c3efd

choropleth maps responsive formik Why I won't be using Edit fiddle React Sync Scroll spark\_streaming Apache Spark with custom\_streaming 18-07 bookmarks All Bookmarks

mflow

Experiments Models

GitHub Docs

mlflow/mflow\_production > Lead\_scoring\_inference\_pipeline/096\_2024\_03\_20\_01

Lead\_scoring\_mlflow\_production096\_2024\_00\_00\_00

Date: 2024-03-28 18:13:58

Source: C:\mflow

User: root

Duration: 9:30

Status: FINISHED

Lifecycle Stage: active

Description

Parameters (24)

Metrics (8)

Tags

Artifacts

models

conda.yaml

conda.yml

python\_env.yaml

requirements.txt

F:\Data\Home\mflow\logs\Lead\_scoring\_training\_pipeline\mlruns\1\2b94292f9def480c8469832dd7c3efd\artifacts\models

2 - registered v1.0.0

Registered on 2024-03-28

MLflow Model

The code snippet below demonstrates 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. Use Mflow docs for how to include input and output schemas with your model. |      |

Make Predictions

Predict on a Spark DataFrame

```
import mlflow
logged_model = 'runs://2b94292f9def480c8469832dd7c3efd/models/'

# Load model as a Spark UDF. Overloads result_type of the model does not require double column.
loaded_model = mlflow.pyfunc.spark_udf(logged_model, mlflow.pyfunc.spark_udf_model_type(logged_model))

# Predict on a Spark DataFrame
columns = ['id','age','income']
df = loaded_model.predict(spark.createDataFrame(loaded_model.columns(), loaded_model.columns()))
```

Predict on a Pandas DataFrame

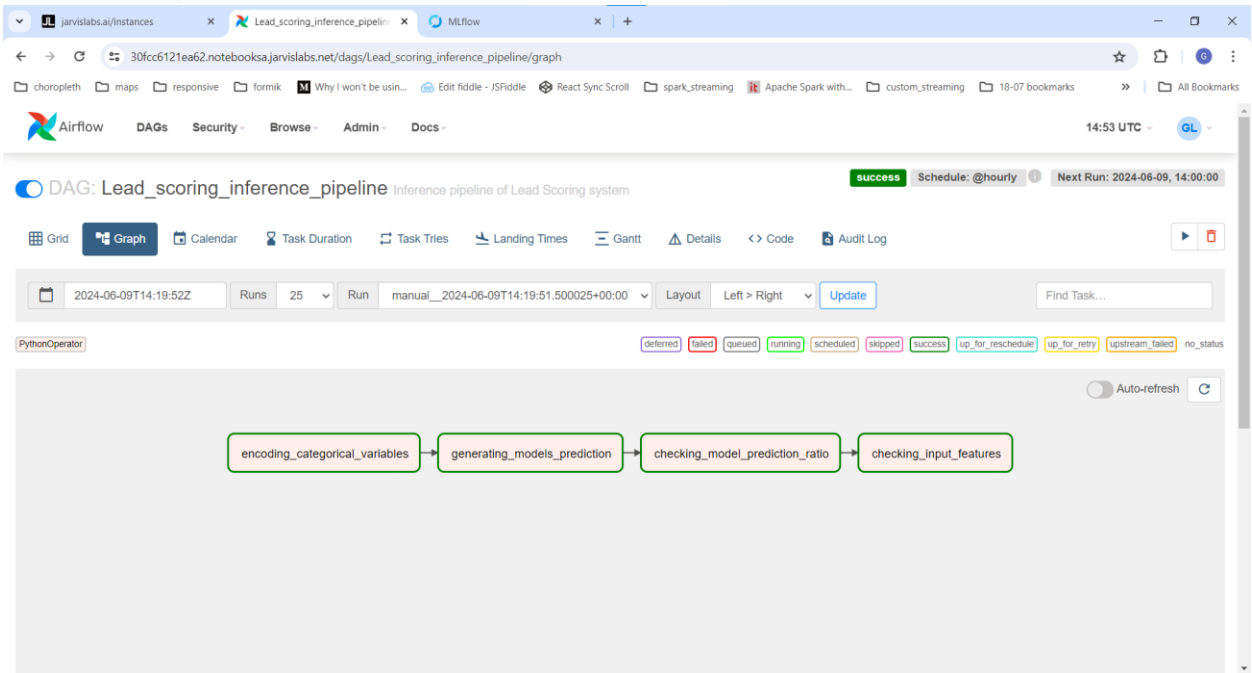
```
import mlflow
logged_model = 'runs://2b94292f9def480c8469832dd7c3efd/models/'

# Load model as a PyFuncModel
loaded_model = mlflow.pyfunc.spark_udf(logged_model)

# Predict on a Pandas DataFrame
import pandas as pd
loaded_model.predict(pd.DataFrame([{"id":1}]))
```

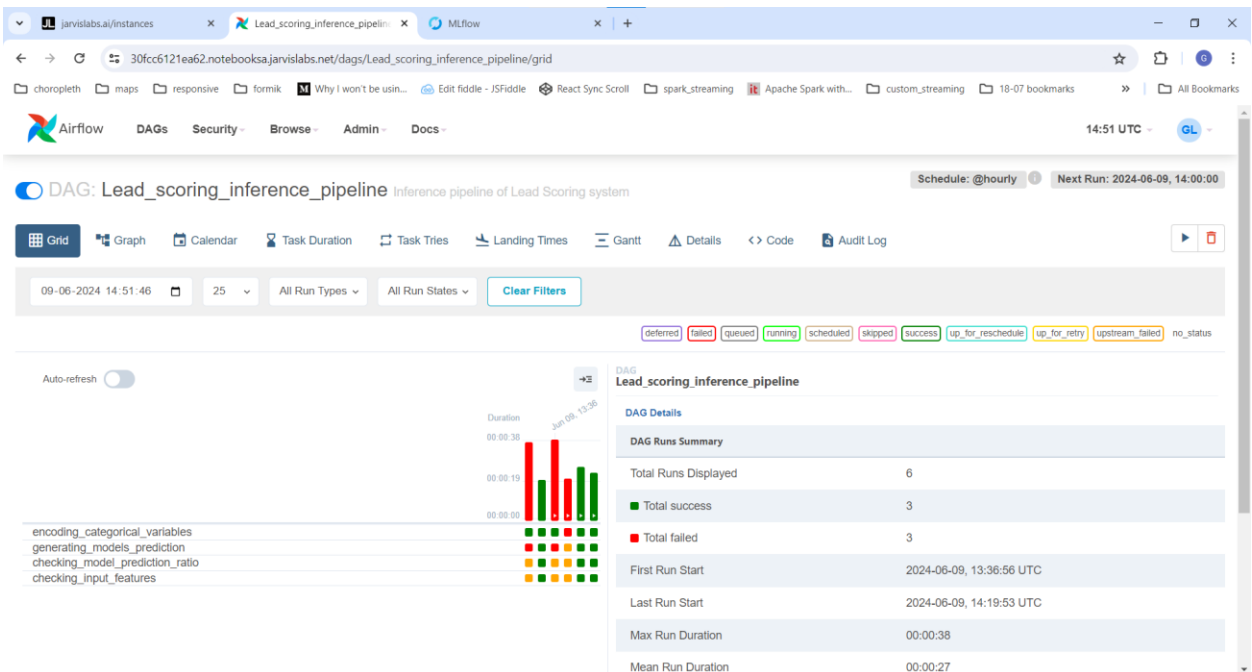
### Inference Pipeline:

## 1. Airflow UI Graph:



The screenshot shows the Airflow UI Graph view for the DAG 'Lead\_scoring\_inference\_pipeline'. The interface includes a top navigation bar with 'Airflow', 'DAGs', 'Security', 'Browse', 'Admin', and 'Docs'. The main header displays the DAG name, its description 'Inference pipeline of Lead Scoring system', and a status 'success' with a schedule of '@hourly' and a next run time of '2024-06-09, 14:00:00'. Below the header, there are tabs for 'Grid', 'Graph', 'Calendar', 'Task Duration', 'Task Tries', 'Landing Times', 'Gantt', 'Details', 'Code', and 'Audit Log'. The 'Graph' tab is selected, showing a workflow with four tasks: 'encoding\_categorical\_variables', 'generating\_models\_prediction', 'checking\_model\_prediction\_ratio', and 'checking\_input\_features'. The tasks are connected by arrows, indicating a sequential flow. A search bar and a 'Find Task...' input are visible. The bottom of the graph shows a list of task states: deferred, failed, queued, running, scheduled, skipped, success, up\_for\_reschedule, up\_for\_retry, upstream\_failed, and no\_status.

## 2. Airflow UI Grid:



The screenshot shows the Airflow UI Grid view for the DAG 'Lead\_scoring\_inference\_pipeline'. The interface includes a top navigation bar with 'Airflow', 'DAGs', 'Security', 'Browse', 'Admin', and 'Docs'. The main header displays the DAG name, its description 'Inference pipeline of Lead Scoring system', and a status 'success' with a schedule of '@hourly' and a next run time of '2024-06-09, 14:00:00'. Below the header, there are tabs for 'Grid', 'Graph', 'Calendar', 'Task Duration', 'Task Tries', 'Landing Times', 'Gantt', 'Details', 'Code', and 'Audit Log'. The 'Grid' tab is selected, showing a list of task runs. The 'Grid' view includes a search bar, a 'Clear Filters' button, and a 'Duration' bar chart. The 'Duration' bar chart shows the duration of task runs for the tasks: 'encoding\_categorical\_variables', 'generating\_models\_prediction', 'checking\_model\_prediction\_ratio', and 'checking\_input\_features'. The 'Grid' view also includes a table of task runs with columns for task name, start time, and duration. The 'DAG Details' section on the right provides a summary of the DAG runs, including the total number of runs displayed (6), total success (3), total failed (3), first run start time (2024-06-09, 13:36:56 UTC), last run start time (2024-06-09, 14:19:53 UTC), max run duration (00:00:38), and mean run duration (00:00:27).

| Task Name                       | Start Time               | Duration |
|---------------------------------|--------------------------|----------|
| encoding_categorical_variables  | 2024-06-09, 13:36:56 UTC | 00:00:38 |
| generating_models_prediction    | 2024-06-09, 13:36:56 UTC | 00:00:11 |
| checking_model_prediction_ratio | 2024-06-09, 13:36:56 UTC | 00:00:01 |
| checking_input_features         | 2024-06-09, 13:36:56 UTC | 00:00:01 |

| Task Name                       | Start Time               | Duration |
|---------------------------------|--------------------------|----------|
| encoding_categorical_variables  | 2024-06-09, 13:36:56 UTC | 00:00:38 |
| generating_models_prediction    | 2024-06-09, 13:36:56 UTC | 00:00:11 |
| checking_model_prediction_ratio | 2024-06-09, 13:36:56 UTC | 00:00:01 |
| checking_input_features         | 2024-06-09, 13:36:56 UTC | 00:00:01 |

| Task Name                       | Start Time               | Duration |
|---------------------------------|--------------------------|----------|
| encoding_categorical_variables  | 2024-06-09, 13:36:56 UTC | 00:00:38 |
| generating_models_prediction    | 2024-06-09, 13:36:56 UTC | 00:00:11 |
| checking_model_prediction_ratio | 2024-06-09, 13:36:56 UTC | 00:00:01 |
| checking_input_features         | 2024-06-09, 13:36:56 UTC | 00:00:01 |