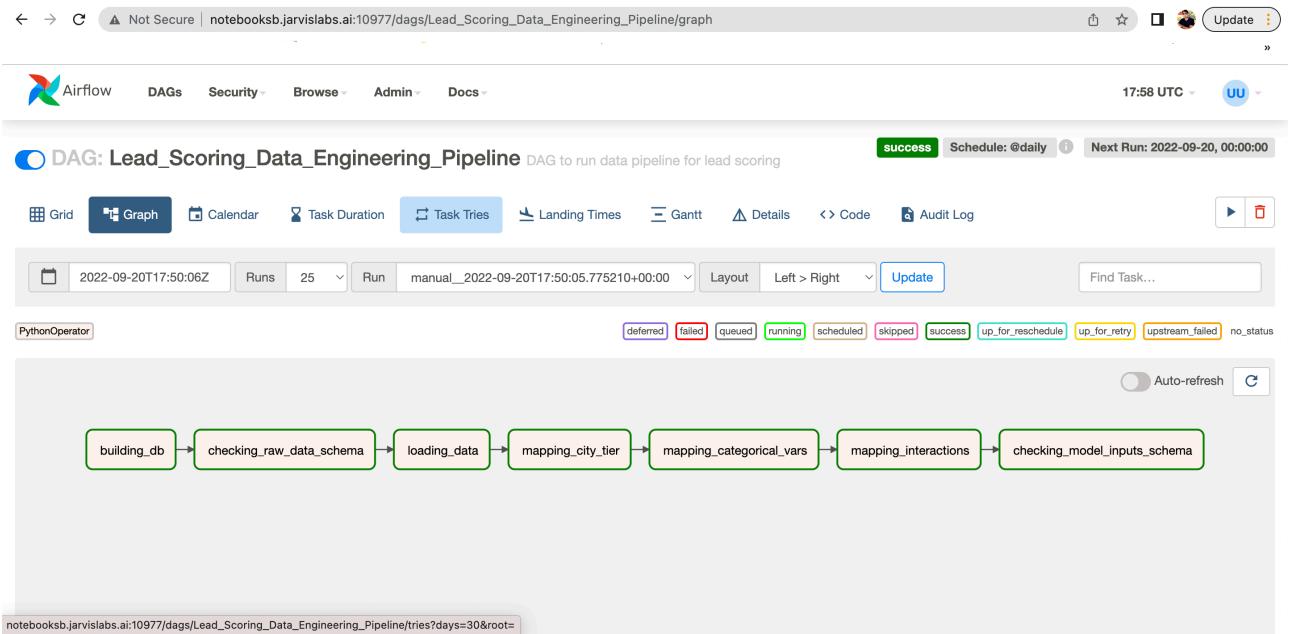


# AIRFLOW HOMEPAGE

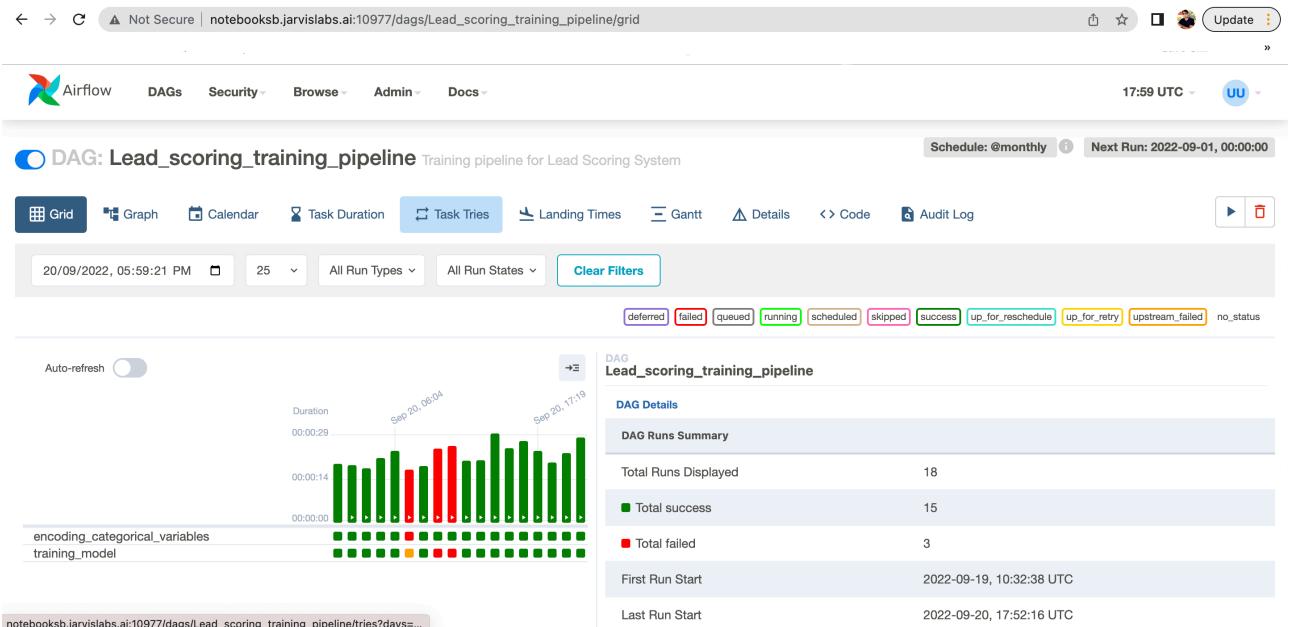
The screenshot shows the Airflow homepage with the title "AIRFLOW HOMEPAGE". At the top, there is a navigation bar with links for "DAGs", "Security", "Browse", "Admin", and "Docs". The main content area is titled "DAGs" and displays a table of active DAGs. The table includes columns for "DAG", "Owner", "Runs", "Schedule", "Last Run", "Next Run", and "Recent Tasks". The "Last Run" column shows the date and time of the most recent run, and the "Next Run" column shows the scheduled date and time. The "Recent Tasks" column shows a grid of task states. A search bar at the top right allows users to search for DAGs by tag or name. At the bottom of the table, there is a pagination control showing page 1 of 3.

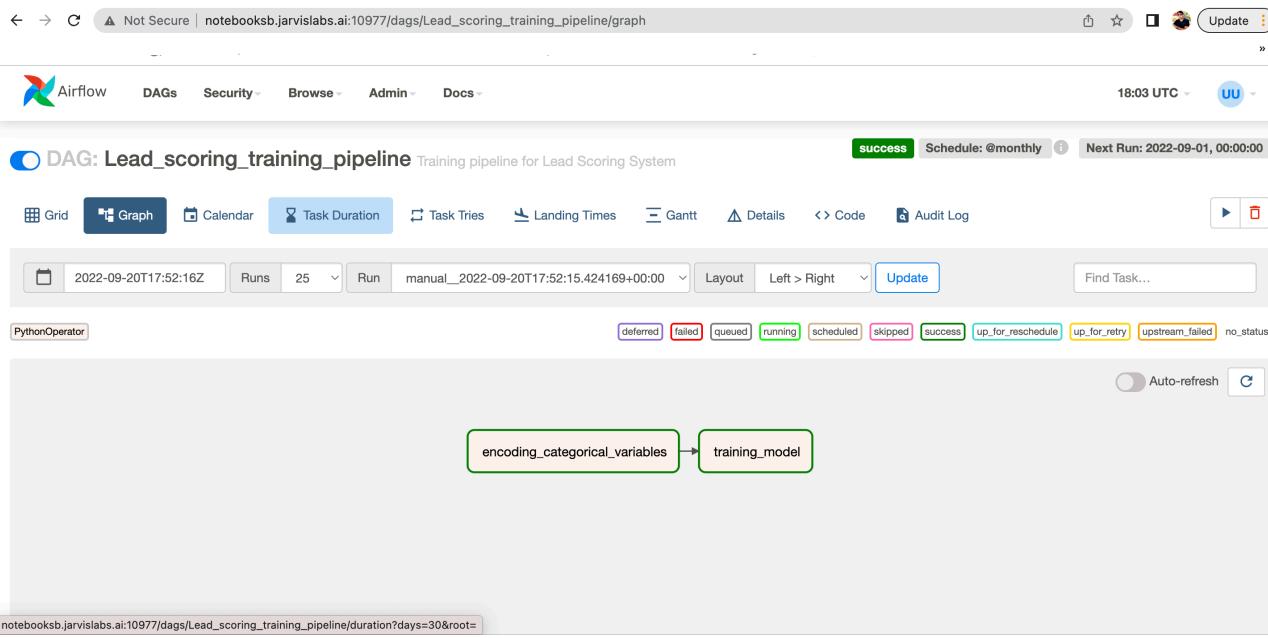
# DATA PIPELINE

The screenshot shows the Data Pipeline interface for the "Lead Scoring Data Engineering Pipeline". The top navigation bar includes links for "DAGs", "Security", "Browse", "Admin", and "Docs". The main content area is titled "DAG: Lead\_Scoring\_Data\_Engineering\_Pipeline" and displays a summary of the DAG's runs. It shows the schedule (@daily) and the next run (2022-09-20, 00:00:00). Below this, there are tabs for "Grid", "Graph", "Calendar", "Task Duration", "Task Tries", "Landing Times", "Gantt", "Details", "Code", and "Audit Log". The "Task Tries" tab is selected. The interface includes filters for "All Run Types" and "All Run States" (with options like deferred, failed, queued, running, scheduled, skipped, success, up\_for\_reschedule, up\_for\_retry, upstream\_failed, and no\_status). A "Clear Filters" button is also present. On the left, a sidebar lists tasks: "building\_db", "checking\_raw\_data\_schema", "loading\_data", "mapping\_city\_tier", and "mapping\_categorical\_vars". The main area features a Gantt chart for the DAG, showing task durations and execution times (e.g., Sep 19, 00:00 to Sep 20, 17:00). A legend indicates task status: green for success, red for failure, and grey for other states. A "DAG Details" section provides a summary of the DAG's runs, including total runs (18), total successes (16), and total failures (2). It also shows the first and last run start times.

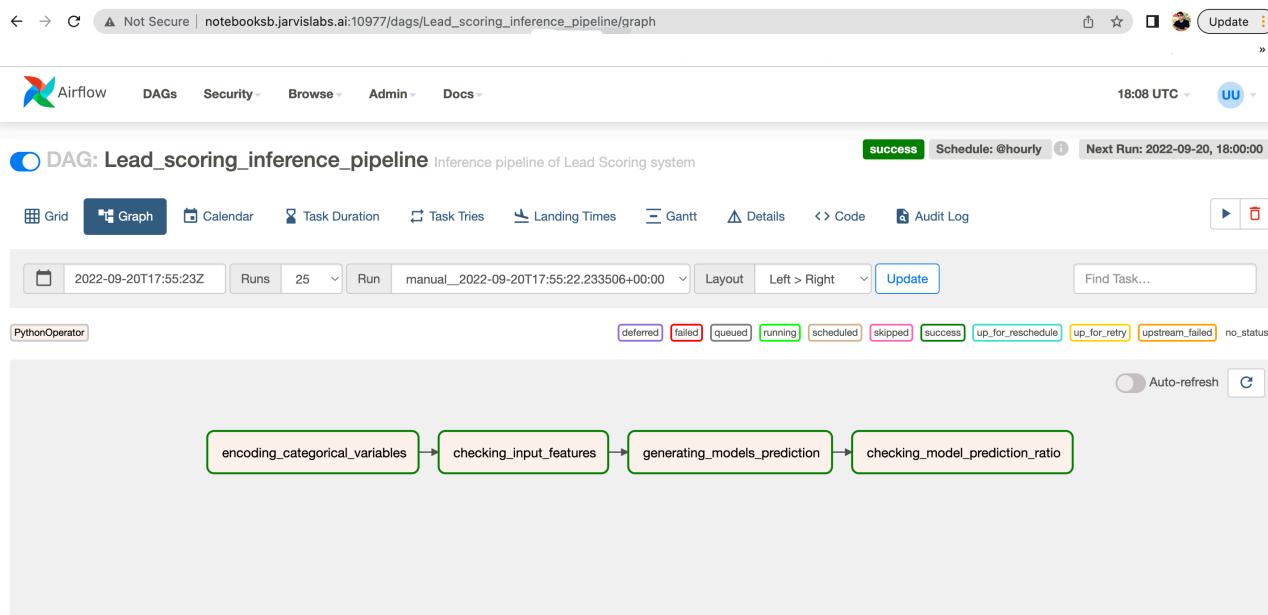


# Training Pipeline

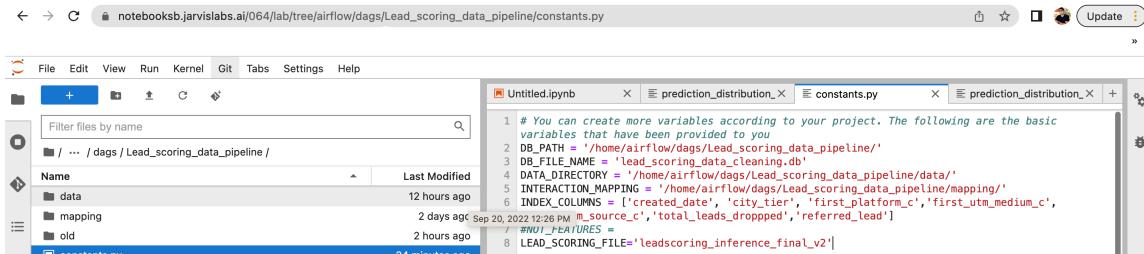




## Inference Pipeline

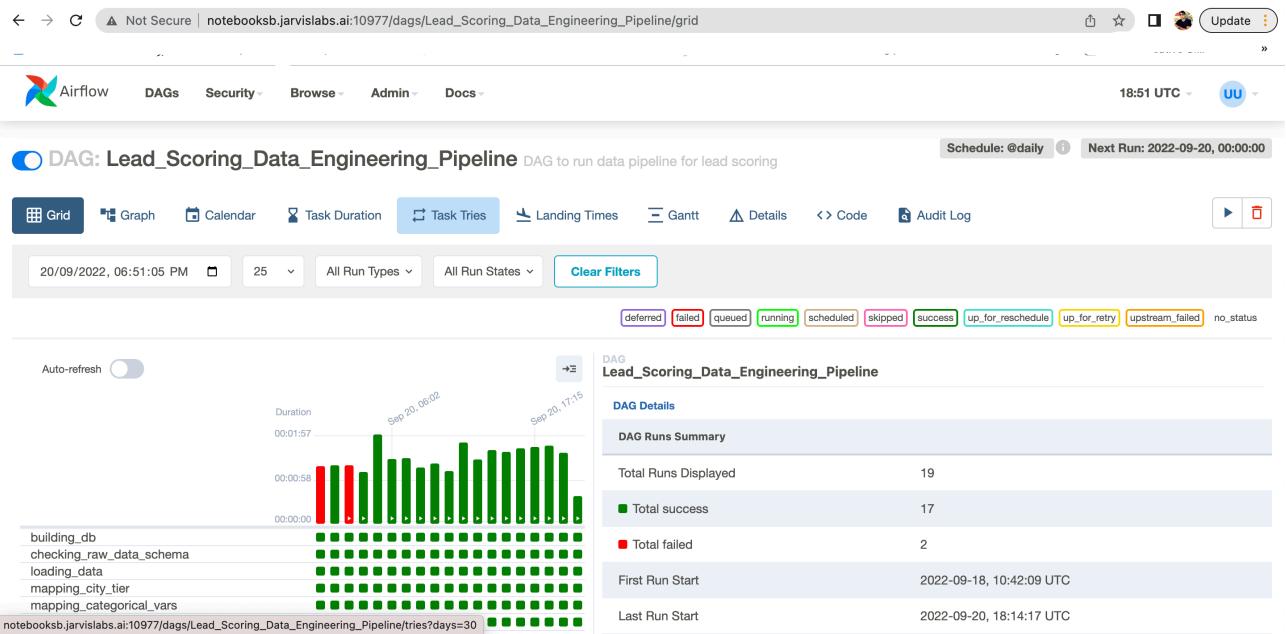


# After making changes in Data Pipeline for Inference file

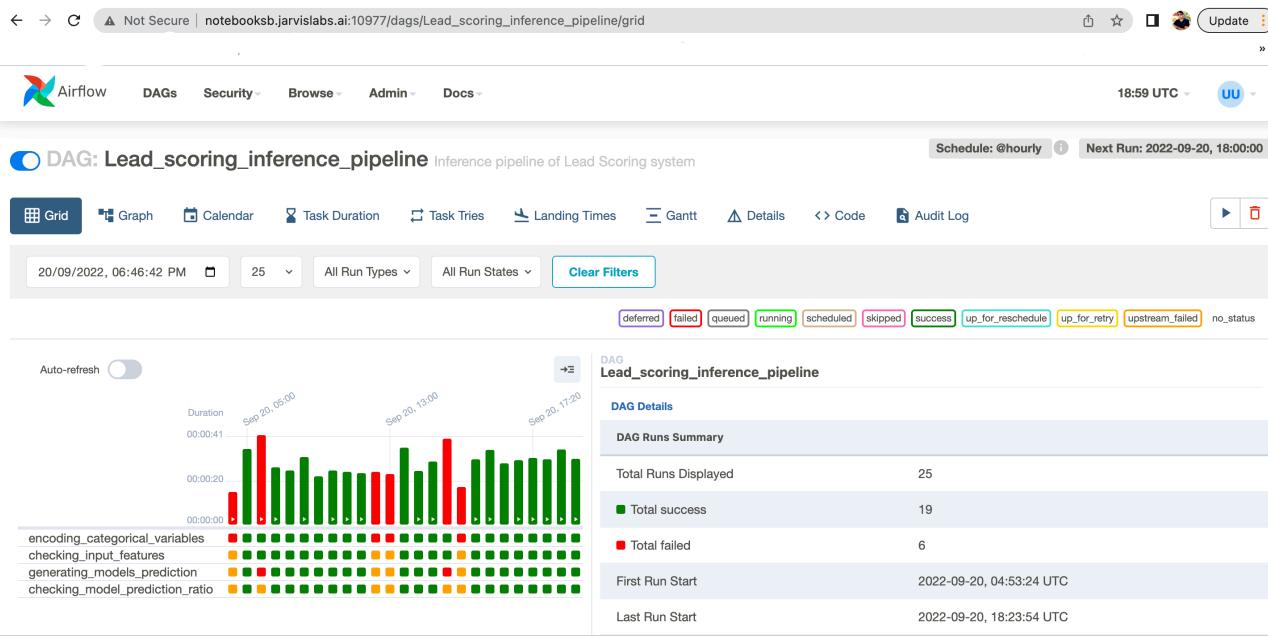


```
# You can create more variables according to your project. The following are the basic variables that have been provided to you
DB_PATH = '/home/airflow/dags/Lead_scoring_data_pipeline/'
TABLES = ['leads', 'building_db', 'mapping']
DATA_DIRECTORY = '/home/airflow/dags/lead_scoring_data_pipeline/data'
INTERACTION_MAPPING = '/home/airflow/dags/lead_scoring_data_pipeline/mapping/'
INDEX_COLUMNS = ['created_date', 'city_tier', 'first_platform_c', 'first_utm_medium_c',
                  'm_source_c', 'total_leads_dropped', 'referred_lead']
# API FEATURES =
LEAD_SCORING_FILE='leadscore_inference_final_v2'
```

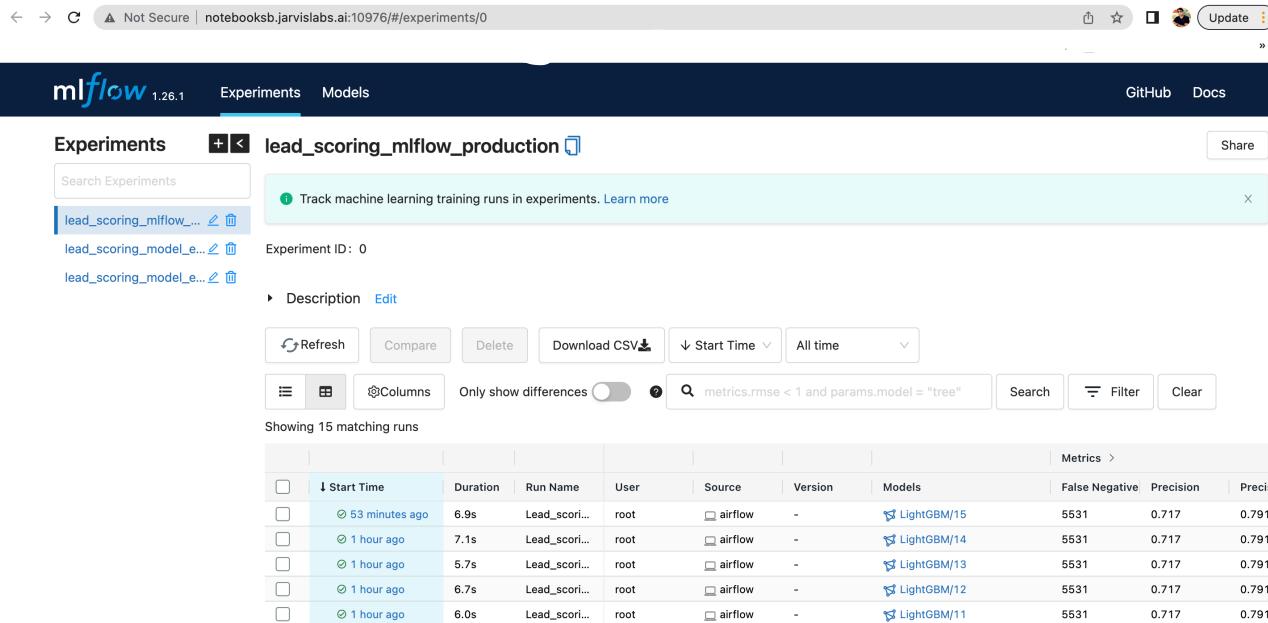
## Data Pipeline



## Data Inference



# ML Flow Server



The screenshot shows the mlflow UI for managing registered models. At the top, there's a header with the mlflow logo (1.26.1), navigation links for Experiments and Models (the latter is selected), and links for GitHub and Docs. Below the header is a banner with a tip about sharing models and a 'Create Model' button. A search bar with a placeholder 'Search by model name' and buttons for 'Search', 'Filter', and 'Clear' is present. The main area displays a table of registered models:

Name	Latest Version	Staging	Production	Last Modified	Tags
LightGBM	Version 15	-	Version 15	2022-09-20 23:25:01	-

Pagination controls at the bottom right show page 1 of 10.

## Models:

The screenshot shows the mlflow UI for a specific model artifact. The URL is /experiments/0/runs/2fc45d56b03244d1882cd9035b8c737f/artifacts/0. On the left, there's a sidebar with 'Tags' and 'Artifacts' sections. The 'models' artifact is selected, showing its contents: MLmodel, conda.yaml, model.pkl, python\_env.yaml, and requirements.txt. The full path is listed as Full Path:/home/Assignment/02\_training\_pipeline/scripts/mlruns/0/2fc45d56b03244d1882cd9035b8c737f/artifacts/0. To the right, there's a detailed view for the 'MLflow Model':

**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 <a href="#">MLflow docs</a> for how to include input and output schema with your model.	

**Make Predictions**  
Predict on a Spark DataFrame:

```
import mlflow
logged_model = 'runs:/2fc45d56b03244d1882cd9035b8c737f/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()

**Predict on a Pandas DataFrame:**

```
import mlflow
logged_model = 'runs:/2fc45d56b03244d1882cd9035b8c737f/models'
```

# Metrics

The screenshot shows the mlflow UI for an experiment named 'Lead\_scoring\_mlflow\_production'. The 'Metrics' section is expanded, displaying 12 metrics with their values:

Name	Value
False Negative	0.531
Precision_0	0.717
Precision_0_0	0.791
Precision_0_1	0.674
Recall_0	0.732
Recall_0_0	0.588
Recall_0_1	0.846
True Negative	20946
F1_0	0.675
F1_1	0.75
roc_auc	0.717
test_accuracy	0.717

Below the metrics, there are sections for 'Tags' and 'Artifacts'. The 'Artifacts' section shows a directory structure for a 'MLflow Model' containing files like 'MLmodel', 'conda.yaml', 'model.pkl', 'parameters.json', and 'requirements.txt'. A note indicates that code snippets demonstrate how to make predictions using the logged model.

# Parameters

The screenshot shows the mlflow UI for the same experiment. The 'Parameters' section is expanded, listing 20 parameters with their values:

Name	Value
boosting_type	gbdt
class_weight	None
criterion	gini
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
objective	None
random_state	42
reg_alpha	0.0
reg_lambda	0.0
silent	warn
subsample	1.0
subsample_for_bin	200000
subsample_freq	0

Below the parameters, there are sections for 'Metrics' (12), 'Tags', and 'Artifacts'.

## Please Note : - Regarding Unit Test and Scripts Files

- Data Pipeline Unit test ipynb file location : Lead\_scoring\_data\_pipeline/AssignmentFolderFiles/scripts/test\_utils\_and\_validation\_check.ipynb
- Data Training pipeline Unit test ipynb file location : Lead\_scoring\_training\_pipeline/AssignmentFolderFiles/Scripts/TestTrainingPipeline.ipynb
- Data Inference pipeline Unit test ipynb file : Lead\_scoring\_inference\_pipeline/AssignmentFolderFiles/scripts/Test\_Inference\_Function.ipynb

All the relevant Scripts files are placed in AssignmentFolder/scripts under respective Pipelines folder