






# Model Experimentation Part



## Screenshot of ml flow during experimentation



Experiments **+** **<** Lead\_scoring\_model\_experimentation 


Search Experiments

Default  

Lead\_scoring\_model...  







Lead\_scoring  





Lead\_scoring\_mlflow...  

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 40 matching runs

								Metrics >
<input type="checkbox"/>	Start Time	Duration	Run Name	User	Source	Version	Models	AUC
<input type="checkbox"/>	2 hours ago		Session Initi...	root	ipykernel...	-	-	-
<input type="checkbox"/>	5 days ago		Session Initi...	root	ipykernel...	-	-	-
<input type="checkbox"/>	9 days ago		Session Initi...	root	ipykernel...	-	-	-
<input type="checkbox"/>	9 days ago		Session Initi...	root	ipykernel...	-	-	-

## Experiment before features dropping

```
# setup pycaret
exp_lead_scoring = setup(data=dataset, target = 'app_complete_flag',
                          fold_shuffle=True,
                          session_id = 42,
                          normalize = True,
                          transformation = True,
                          remove_multicollinearity = True, multicollinearity_threshold = 0.95,
                          n_jobs=-1, use_gpu=True,
                          log_experiment=True, experiment_name='Lead_scoring_model_experimentation',
                          log_plots=True, log_data=True,
                          silent=True, verbose=True,
                          log_profile=False)
```

	Description	Value
0	session_id	42
1	Target	app_complete_flag
2	Target Type	Binary
3	Label Encoded	None
4	Original Data	(238964, 12)
5	Missing Values	False

(ipykernel) | Idle Mode: Command Ln 1, Col 1 lead\_scoring\_model\_experimentation.ipynb

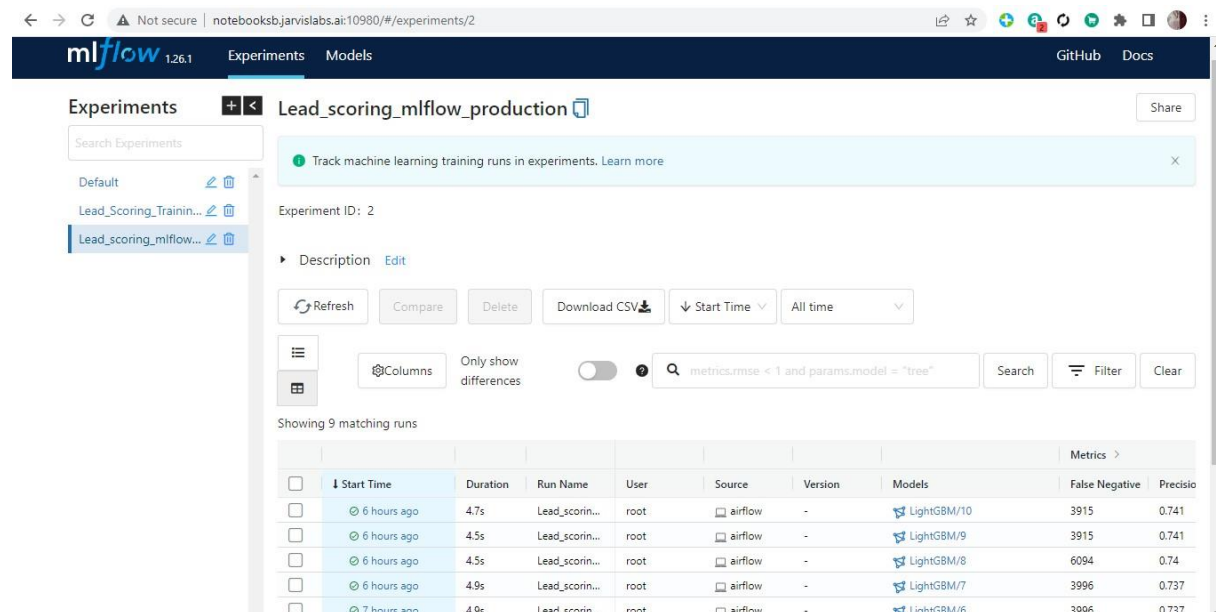
# Experiment setup after features dropping

[14]: # setup pycaret  
from pycaret.classification import \*  
  
exp\_lead\_scoring = setup(data=dataset\_cols, target = 'app\_complete\_flag',  
remove\_multicollinearity = True, multicollinearity\_threshold = 0.95,  
categorical\_features = ['city\_tier', 'first\_platform\_c', 'first\_utm\_medium\_c', 'first\_utm\_source\_c',  
fold\_shuffle=True,  
session\_id = 42,  
n\_jobs=-1,use\_gpu=True,  
log\_experiment=True,experiment\_name='Lead\_scoring',  
log\_plots=True, log\_data=True,  
silent=True, verbose=True,  
log\_profile=False)

	Description	Value
0	session_id	42
1	Target	app_complete_flag
2	Target Type	Binary
3	Label Encoded	None
4	Original Data	(238964, 7)
5	Missing Values	False
6	Numeric Features	1

## Model Training Part

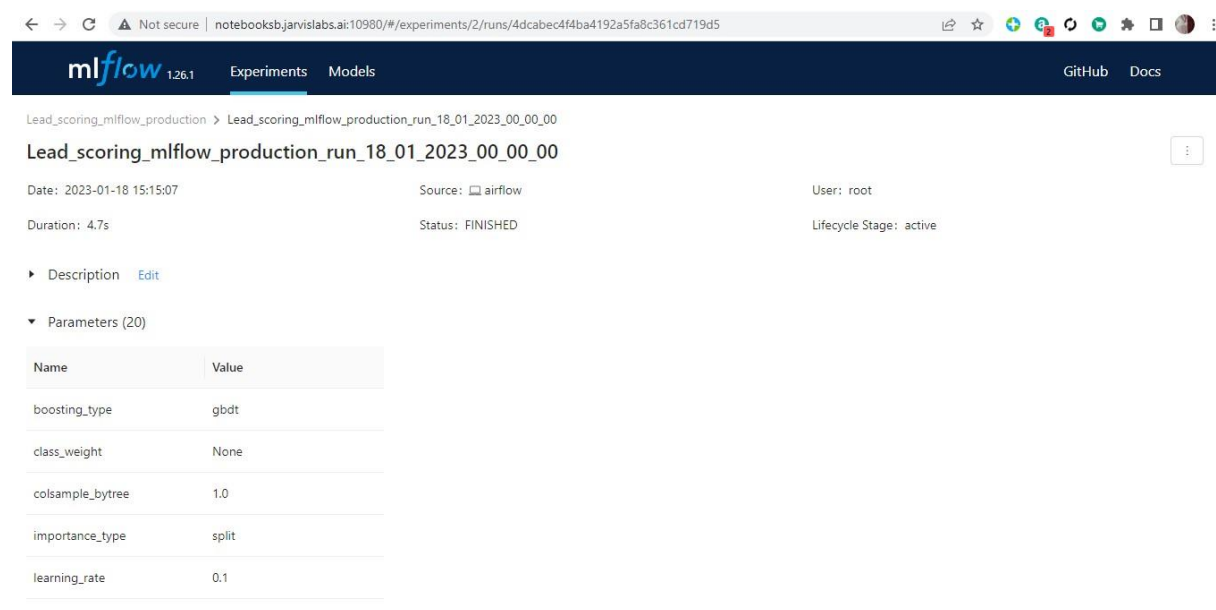
## After Running multiple times mlflow\_production



The screenshot shows the mlflow Experiments page for the experiment 'Lead\_scoring\_mlflow\_production'. The interface includes a sidebar with a search bar and a list of experiments. The main panel displays the experiment details, including the Experiment ID (2), a description, and a table of runs. The table shows 9 matching runs, with columns for Start Time, Duration, Run Name, User, Source, Version, Models, False Negative, and Precision. The runs are sorted by Start Time, showing runs from 6 hours ago to 7 hours ago. The models listed are LightGBM/10, LightGBM/9, LightGBM/8, LightGBM/7, and LightGBM/6.

	Start Time	Duration	Run Name	User	Source	Version	Models	False Negative	Precision
<input type="checkbox"/>	6 hours ago	4.7s	Lead_scoring...	root	airflow	-	LightGBM/10	3915	0.741
<input type="checkbox"/>	6 hours ago	4.5s	Lead_scoring...	root	airflow	-	LightGBM/9	3915	0.741
<input type="checkbox"/>	6 hours ago	4.5s	Lead_scoring...	root	airflow	-	LightGBM/8	6094	0.74
<input type="checkbox"/>	6 hours ago	4.9s	Lead_scoring...	root	airflow	-	LightGBM/7	3996	0.737
<input type="checkbox"/>	7 hours ago	4.9s	Lead_scoring...	root	airflow	-	LightGBM/6	3996	0.737

Below some screenshot show details of LightGBM/10 model from above mlflow production



The screenshot shows the mlflow Models page for the model 'Lead\_scoring\_mlflow\_production\_run\_18\_01\_2023\_00\_00\_00'. The interface includes a sidebar with a search bar and a list of models. The main panel displays the model details, including the Date (2023-01-18 15:15:07), Source (airflow), User (root), Duration (4.7s), Status (FINISHED), and Lifecycle Stage (active). The Parameters section shows a table of parameters and their values.

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
objective	None
random_state	42
reg_alpha	0.0
reg_lambda	0.0
silent	warn
subsample	1.0
subsample_for_bin	200000

Metrics (13)	
Name	Value
False Negative	3915
Precision	0.741
Precision_0	0.795
Precision_1	0.706
Recall	0.751
Recall_0	0.643
Recall_1	0.838
True Negative	15199
f1 score	0.739
f1_0	0.711
f1_1	0.766
roc_auc	0.751
test_accuracy	0.742

▼ Artifacts

models

MLmodel  
conda.yaml  
model.pkl  
python\_env.yaml  
requirements.txt

Full Path: /home/Assignment/mlruns/2/4dcabec4f4ba4192a5fa8c361cd719d5/artifacts/models

LightGBM, v10  
Registered on 2023/01/18

### 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:/4dcabec4f4ba4192a5fa8c361cd719d5/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:/4dcabec4f4ba4192a5fa8c361cd719d5/models'

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

# Predict on a Pandas DataFrame.
import pandas as pd
loaded_model.predict(pd.DataFrame(data))
```

← → ↻ ⚠ Not secure | notebooks.jarvislabs.ai:10980/#/models/LightGBM/versions/10 🔍 📄 ⭐ 🌐 🔄 🗑 📁 👤 ⋮

mlflow 1.26.1

Experiments Models

GitHub Docs

Registered Models > LightGBM > Version 10

### Version 10

Registered At: 2023-01-18 15:15:12 Stage: Production Last Modified: 2023-01-18 15:16:22

Source Run: [Lead\\_scoring\\_mlflow\\_production\\_run\\_18\\_01\\_2023\\_00\\_00\\_00](#)

► Description [Edit](#)

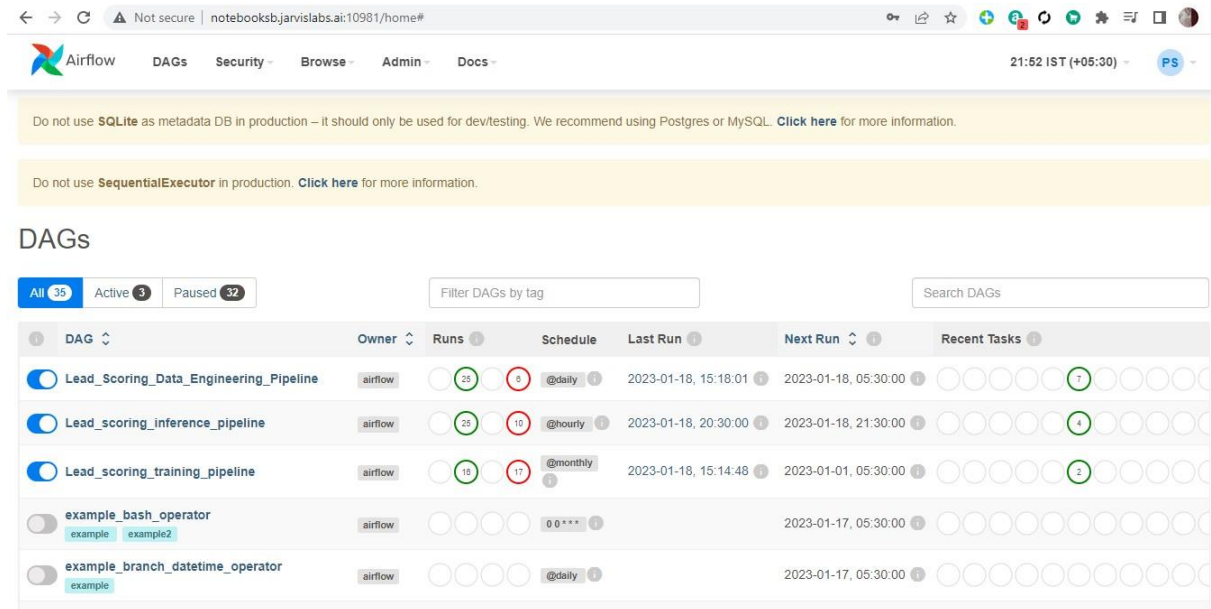
► Tags

▼ Schema

Name	Type
No schema. See <a href="#">MLflow docs</a> for how to include input and output schema with your model.	

# Airflow Part Screenshot

## Airflow Home page



The screenshot shows the Airflow Home page in a web browser. The browser address bar shows the URL `notebooksb.jarvislabs.ai:10981/home#`. The Airflow logo is in the top left, and navigation links for DAGs, Security, Browse, Admin, and Docs are in the top center. The top right shows the time 21:52 IST (+05:30) and a user profile icon labeled PS.

Two yellow warning banners are displayed:

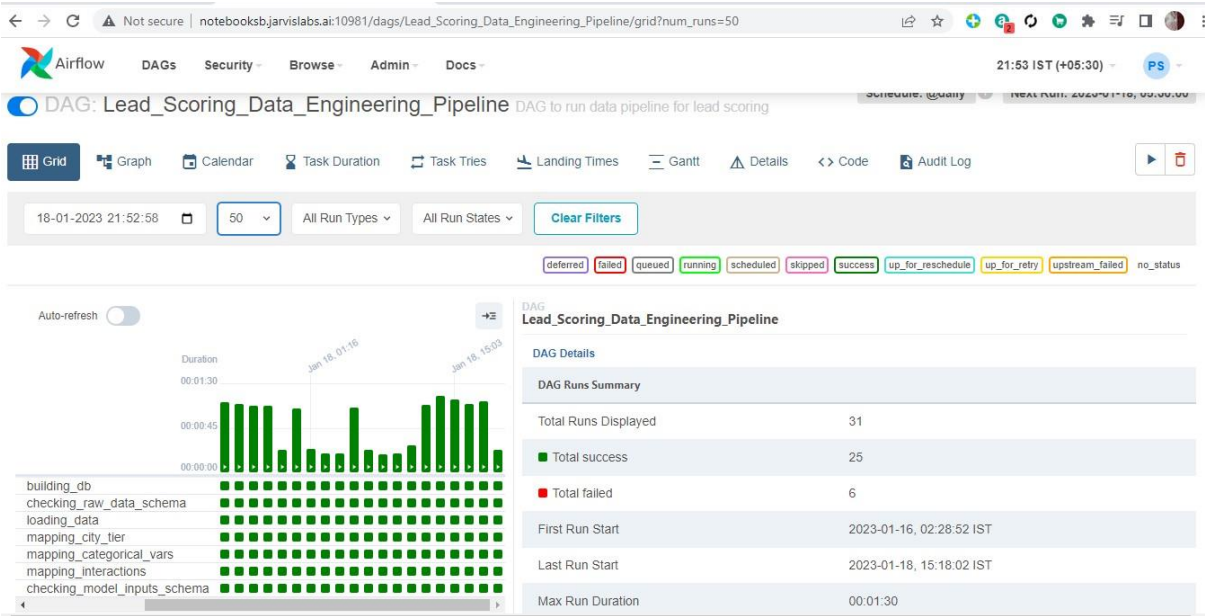
- Do not use **SQLite** as metadata DB in production – it should only be used for dev/testing. We recommend using Postgres or MySQL. [Click here](#) for more information.
- Do not use **SequentialExecutor** in production. [Click here](#) for more information.

### DAGs

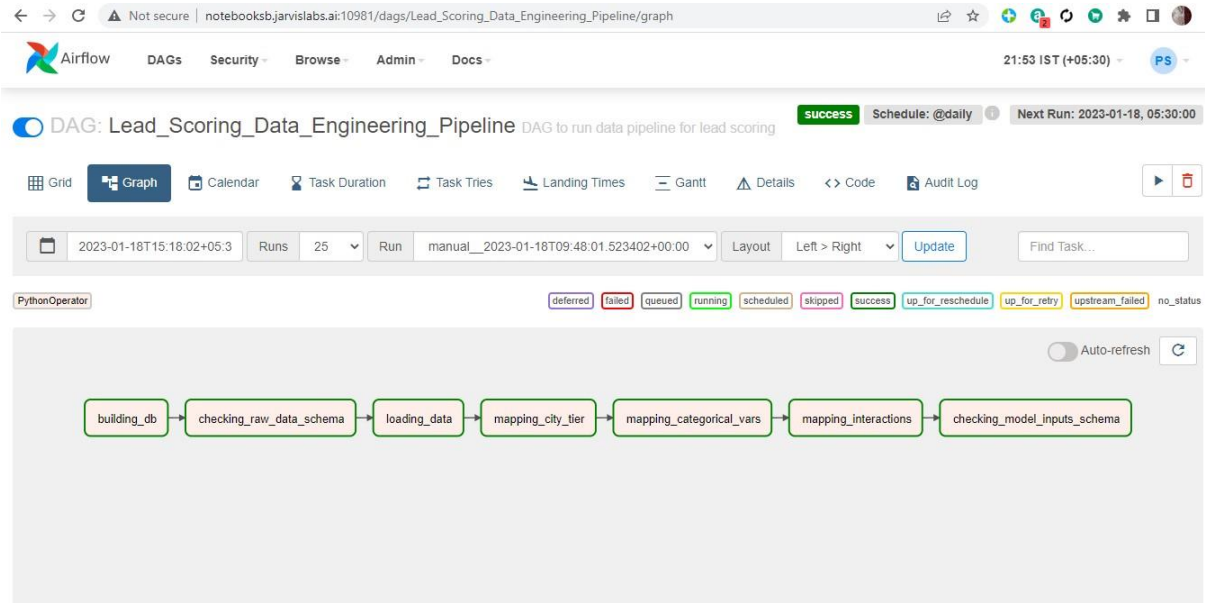
Filters: All (35) Active (3) Paused (32) Filter DAGs by tag Search DAGs

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks
<input checked="" type="checkbox"/> Lead_Scoring_Data_Engineering_Pipeline	airflow	25 (10 failed)	@daily	2023-01-18, 15:18:01	2023-01-18, 05:30:00	7 tasks
<input checked="" type="checkbox"/> Lead_scoring_inference_pipeline	airflow	25 (10 failed)	@hourly	2023-01-18, 20:30:00	2023-01-18, 21:30:00	4 tasks
<input checked="" type="checkbox"/> Lead_scoring_training_pipeline	airflow	15 (17 failed)	@monthly	2023-01-18, 15:14:48	2023-01-01, 05:30:00	2 tasks
<input type="checkbox"/> example_bash_operator example example2	airflow	0 (0 failed)	0 0 * * *		2023-01-17, 05:30:00	0 tasks
<input type="checkbox"/> example_branch_datetime_operator example	airflow	0 (0 failed)	@daily		2023-01-17, 05:30:00	0 tasks

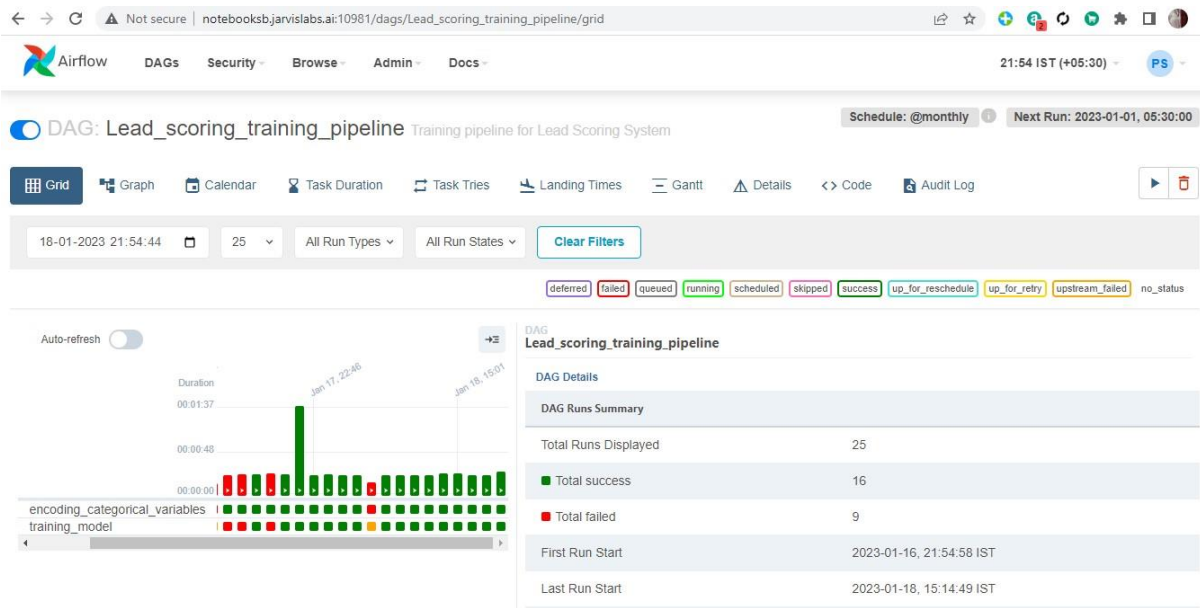
# Data Engineering pipeline Grid view



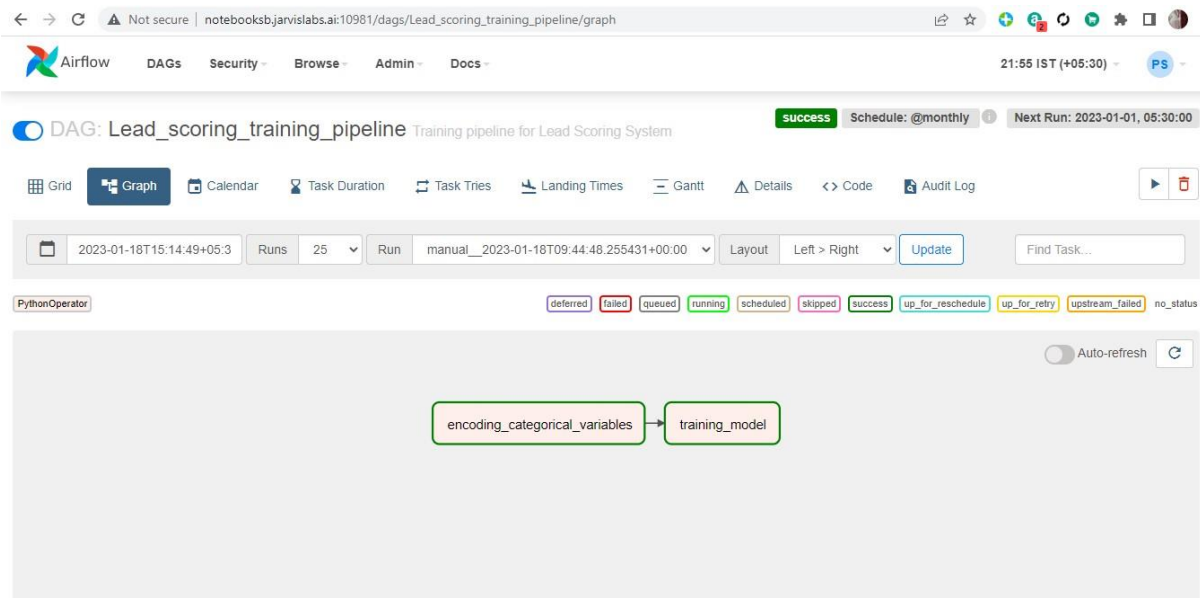
# Data Engineering pipeline Graph view



# Model Training pipeline Grid view

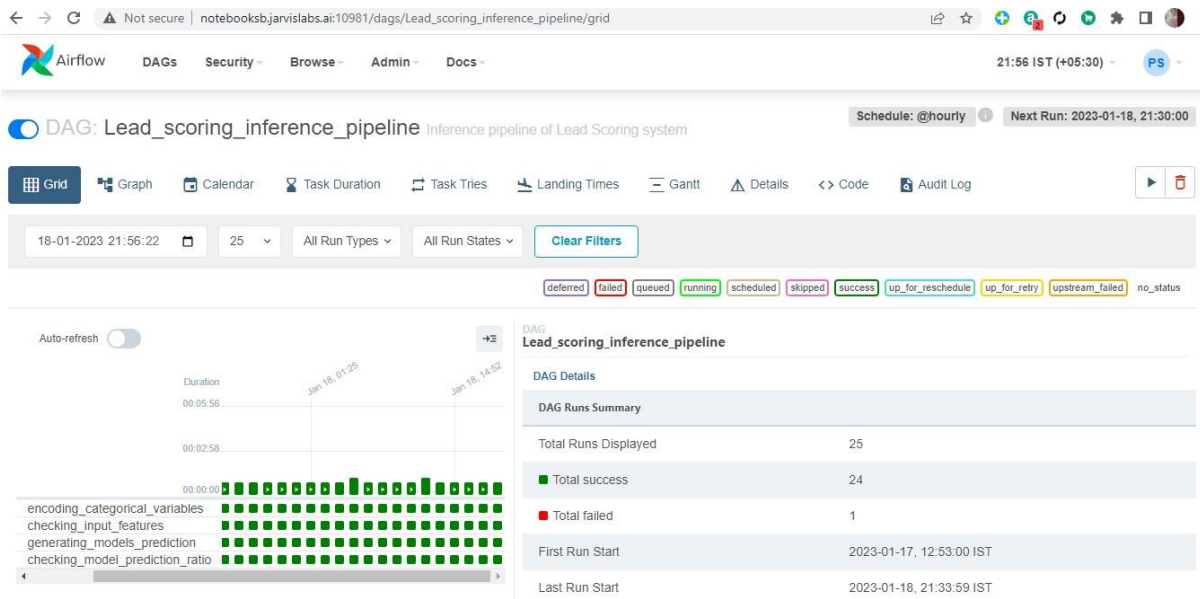


# Model Training pipeline Graph view

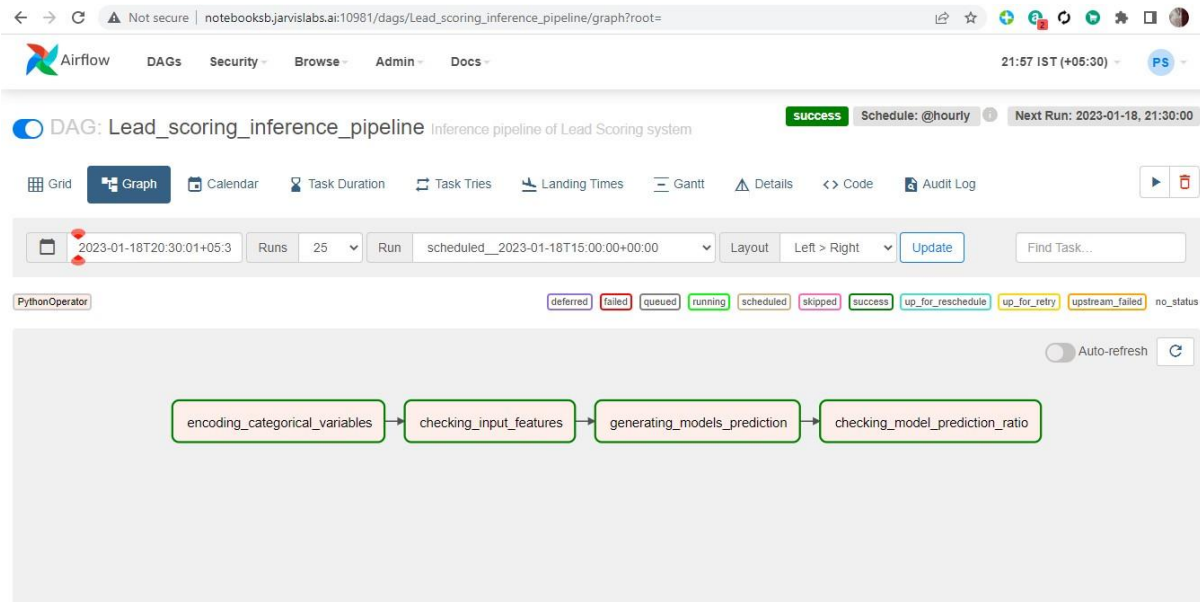




# Inference pipeline Grid view

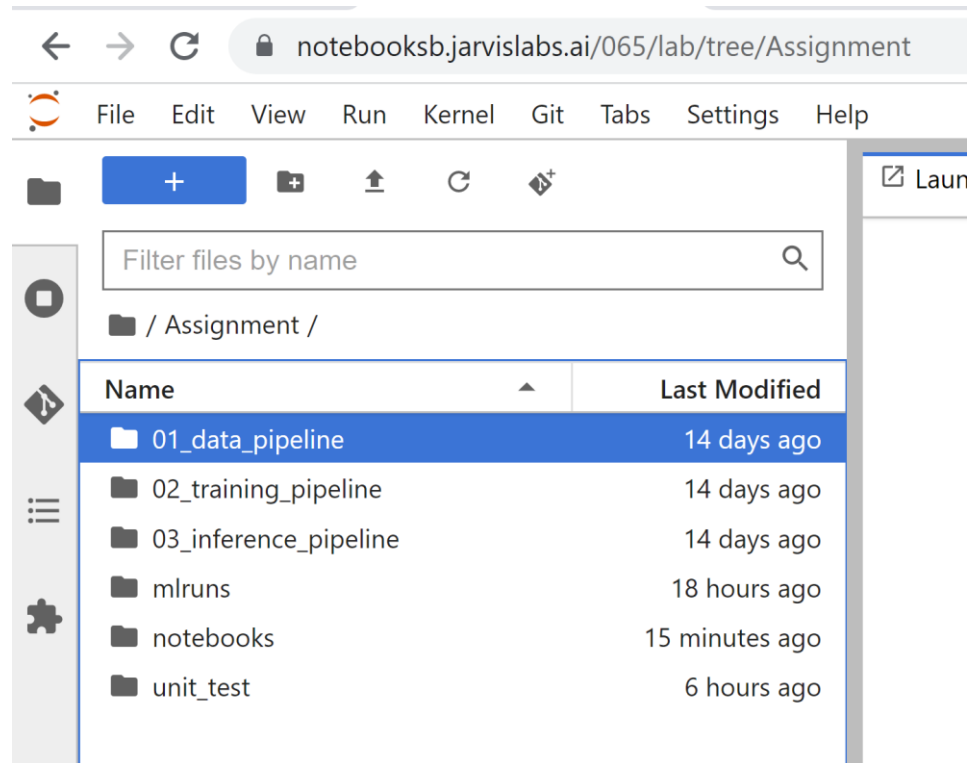


# Inference pipeline Graph view



## Jarvislab Notebook Location Screenshot

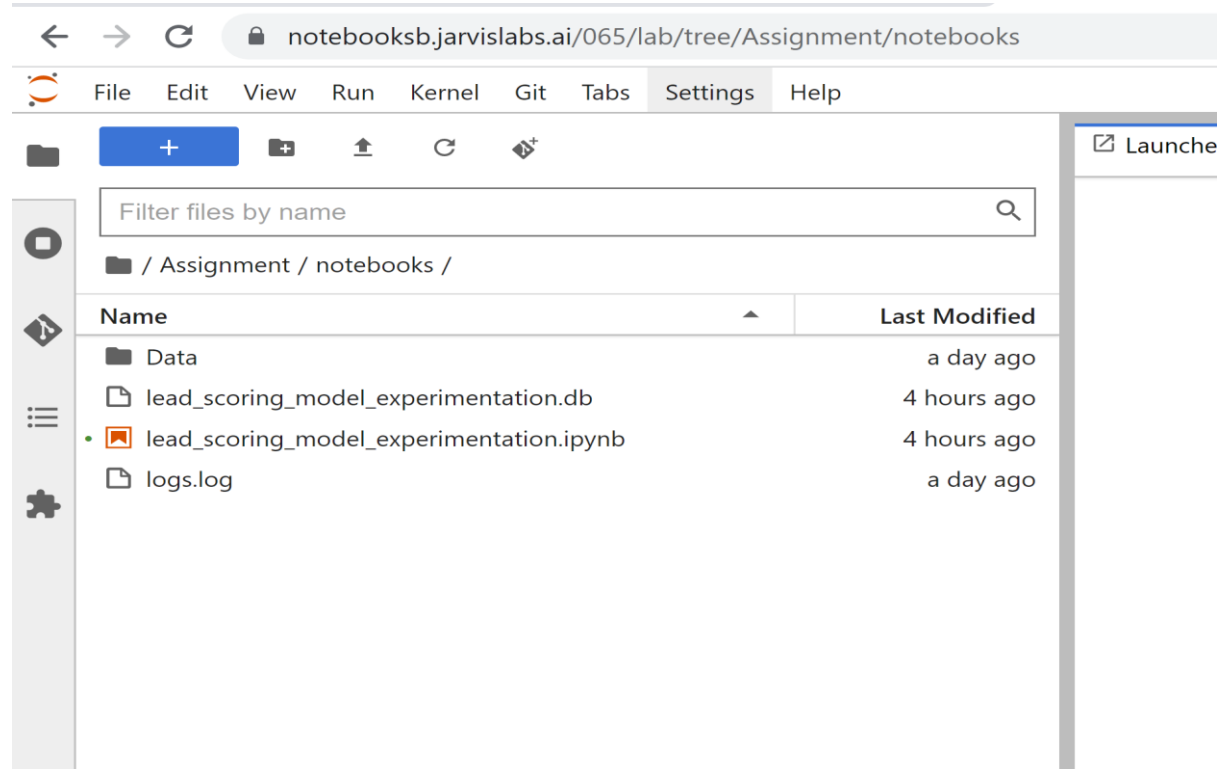
### Assignment directory



The screenshot shows the Jarvislab Notebook interface for the 'Assignment' directory. The browser address bar displays 'notebooksb.jarvislabs.ai/065/lab/tree/Assignment'. The interface includes a menu bar with 'File', 'Edit', 'View', 'Run', 'Kernel', 'Git', 'Tabs', 'Settings', and 'Help'. Below the menu bar is a toolbar with icons for creating new files, uploading, refreshing, and deleting. A search bar labeled 'Filter files by name' is present. The file list shows the current directory path as '/ Assignment /'. The files and folders are listed in a table with columns 'Name' and 'Last Modified'.

Name	Last Modified
01_data_pipeline	14 days ago
02_training_pipeline	14 days ago
03_inference_pipeline	14 days ago
mlruns	18 hours ago
notebooks	15 minutes ago
unit_test	6 hours ago

### Notebooks directory



The screenshot shows the Jarvislab Notebook interface for the 'notebooks' directory. The browser address bar displays 'notebooksb.jarvislabs.ai/065/lab/tree/Assignment/notebooks'. The interface includes a menu bar with 'File', 'Edit', 'View', 'Run', 'Kernel', 'Git', 'Tabs', 'Settings', and 'Help'. Below the menu bar is a toolbar with icons for creating new files, uploading, refreshing, and deleting. A search bar labeled 'Filter files by name' is present. The file list shows the current directory path as '/ Assignment / notebooks /'. The files and folders are listed in a table with columns 'Name' and 'Last Modified'.

Name	Last Modified
Data	a day ago
lead_scoring_model_experimentation.db	4 hours ago
• lead_scoring_model_experimentation.ipynb	4 hours ago
logs.log	a day ago

# Unit\_test directory

←

→

↺

notebooksb.jarvislabs.ai/065/lab/tree/Assignment/unit\_test

File

Edit

View

Run

Kernel

Git

Tabs

Settings

Help

+

Filter files by name

/ Assignment / unit\_test /

Name	Last Modified
city_tier_mapping.py	3 days ago
constants.py	3 days ago
data_validation_checks.py	3 days ago
dummy.ipynb	6 hours ago
interaction_mapping.csv	3 days ago
lead_scoring_data_pipeline.py	3 days ago
leadscoring_test.csv	3 days ago
schema.py	3 days ago
significant_categorical_level.py	3 days ago
test_with_pytest.py	3 days ago
unit_test_cases.db	3 days ago
utils_output.db	3 days ago
utils.py	3 days ago

Launcher

## Airflow/dags directory

The screenshot shows the JupyterLab interface with the URL `notebooksb.jarvislabs.ai/065/lab/tree/airflow/dags`. The top menu bar includes File, Edit, View, Run, Kernel, Git, Tabs, Settings, and Help. Below the menu, there are icons for file operations: a folder, a plus sign, a folder with a plus, an upload arrow, a refresh arrow, and a delete icon. A search bar labeled "Filter files by name" is present. The left sidebar shows a file tree with the path `/ airflow / dags /`. The main panel displays a table of DAGs:

Name	Last Modified
Lead_scoring_data_pipeline	4 hours ago
Lead_scoring_inference_pipeline	28 minutes ago
Lead_scoring_training_pipeline	an hour ago

The screenshot shows the JupyterLab interface with the URL `notebooksb.jarvislabs.ai/065/lab/tree/airflow/dags/Lead_scoring_data_pipeline`. The top menu bar includes File, Edit, View, Run, Kernel, Git, Tabs, Settings, and Help. Below the menu, there are icons for file operations: a folder, a plus sign, a folder with a plus, an upload arrow, a refresh arrow, and a delete icon. A search bar labeled "Filter files by name" is present. The left sidebar shows a file tree with the path `/ ... / dags / Lead_scoring_data_pipeline /`. The main panel displays a table of files and folders within the DAG directory:

Name	Last Modified
data	3 days ago
mapping	3 days ago
constants.py	8 hours ago
data_validation_checks.py	3 days ago
lead_scoring_data_cleaning.db	an hour ago
lead_scoring_data_pipeline.py	3 days ago
schema.py	3 days ago
utils.py	3 days ago

← → ↻ 🔒 notebooks.jarvislabs.ai/065/lab/tree/airflow/dags/Lead\_scoring\_training\_pipeline

File Edit View Run Kernel Git Tabs Settings Help

Filter files by name 🔍

/ ... / dags / Lead\_scoring\_training\_pipeline /

Name	Last Modified
mlruns	2 days ago
constants.py	a day ago
Lead_scoring_mlflow_production.db	8 hours ago
lead_scoring_training_pipeline.py	2 days ago
utils.py	4 hours ago

airf

← → ↻ 🔒 notebooks.jarvislabs.ai/065/lab/tree/airflow/dags/Lead\_scoring\_inference\_pipeline

File Edit View Run Kernel Git Tabs Settings Help

Filter files by name 🔍

/ ... / dags / Lead\_scoring\_inference\_pipeline /

Name	Last Modified
constants.py	4 hours ago
lead_scoring_inference_pipeline.py	a day ago
prediction_distribution.txt	2 hours ago
utils.py	4 hours ago

airflc