

MLOps Assignment

Contents

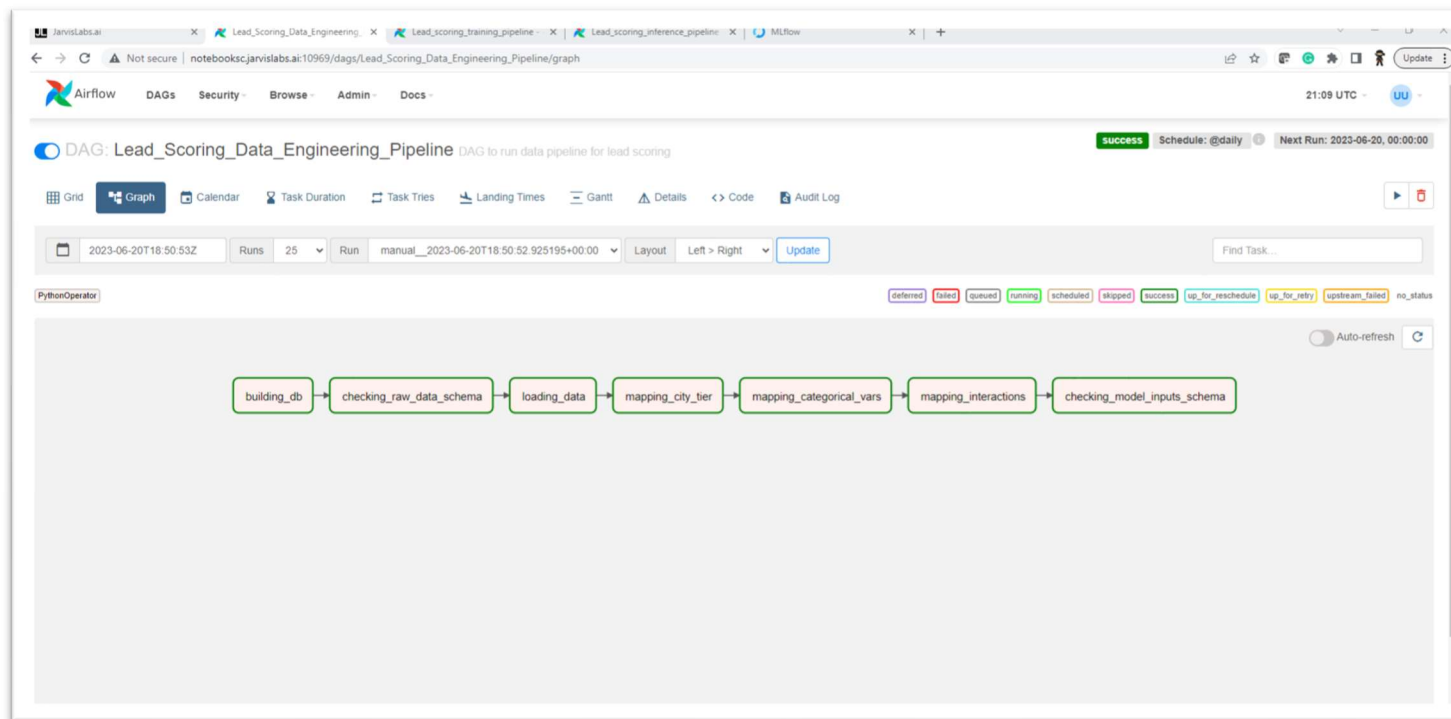
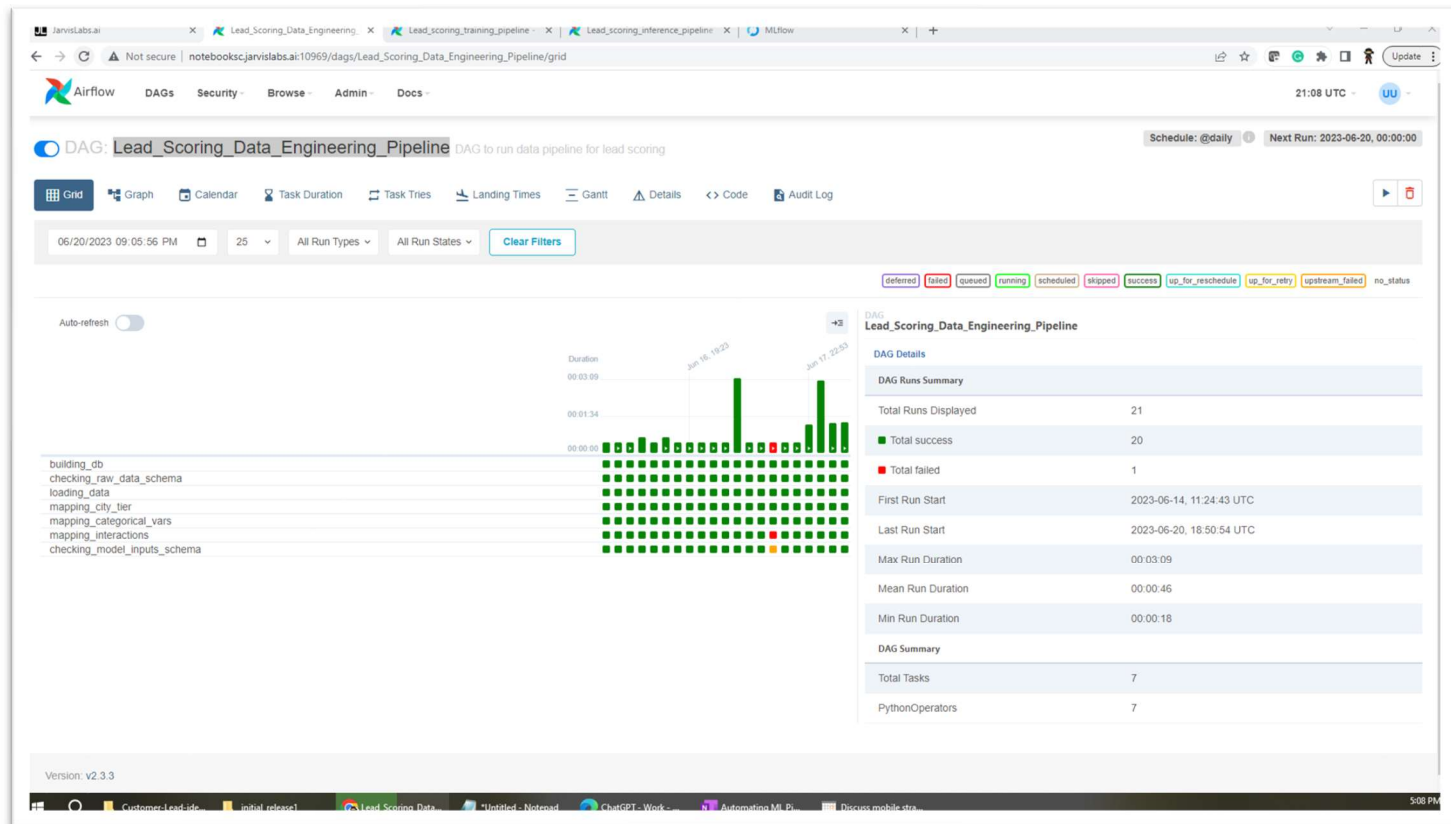
MLOps Assignment	1
1. Initial DAG Landing Page	1
2. Lead_Scoring_Data_Engineering_Pipeline	2
3. Lead_Scoring_Data_Engineering_Pipeline	3
4. Lead_Scoring_Data_Engineering_Pipeline	4
5. Lead_Scoring_Data_Engineering_Pipeline_test	5
6. mlFlow Lead Scoring Model Experimentation	6

1. Initial DAG Landing Page

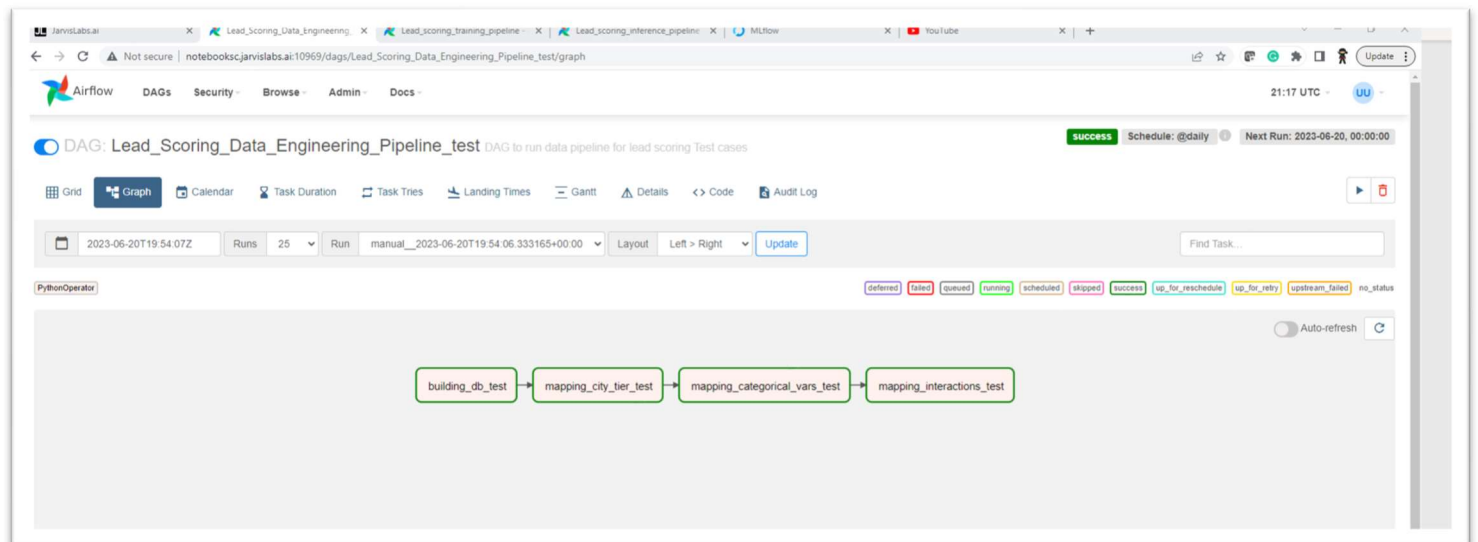
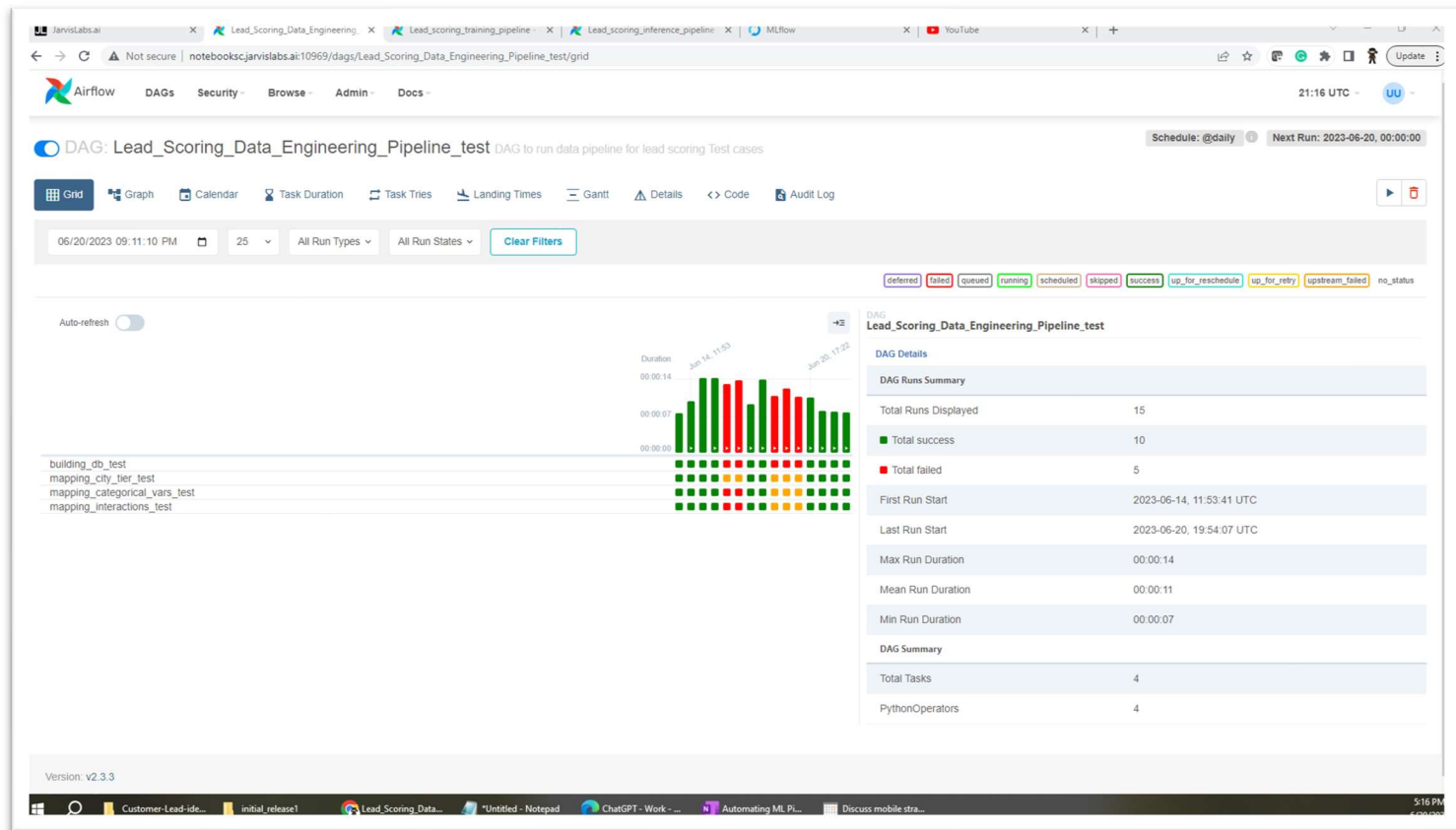
The screenshot displays the Apache Airflow web interface. At the top, there's a navigation bar with links for DAGs, Security, Browse, Admin, and Docs. Below this, a warning message states: "Do not use SQLite as metadata DB in production – it should only be used for dev/testing. We recommend using Postgres or MySQL. Click here for more information." Another warning follows: "Do not use SequentialExecutor in production. Click here for more information." The main section is titled "DAGs" and features a filter for "Filter DAGs by tag" and a search bar. A table lists the DAGs with columns: DAG, Owner, Runs, Schedule, Last Run, Next Run, Recent Tasks, Actions, and Links. The first four DAGs are 'Lead_Scoring_Data_Engineering_Pipeline', 'Lead_Scoring_Data_Engineering_Pipeline_test', 'Lead_scoring_inference_pipeline', and 'Lead_scoring_training_pipeline'. Below these are several example DAGs like 'example_bash_operator', 'example_branch_datetime_operator', 'example_branch_datetime_operator_2', 'example_branch_dop_operator_v3', 'example_branch_labels', 'example_branch_operator', and 'example_branch_python_operator_decorator'. The bottom of the page shows a taskbar with various applications open, including 'Customer-Lead-ide...', 'initial_release1', 'DAGs - Airflow - G...', 'Untitled - Notepad', 'ChatGPT - Work - ...', and 'General (GlobusOn...'. The system clock indicates 2:38 PM on 6/20/2023.

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
Lead_Scoring_Data_Engineering_Pipeline	airflow	19	@daily	2023-06-20, 17:05:14	2023-06-20, 00:00:00	7	[Run] [Cancel] [More]	
Lead_Scoring_Data_Engineering_Pipeline_test	airflow	5	@daily	2023-06-20, 18:11:38	2023-06-20, 00:00:00	4	[Run] [Cancel] [More]	
Lead_scoring_inference_pipeline	airflow	15	@hourly	2023-06-20, 18:13:44	2023-06-20, 18:00:00	4	[Run] [Cancel] [More]	
Lead_scoring_training_pipeline	airflow	15	@monthly	2023-06-20, 18:17:47	2023-06-01, 00:00:00	2	[Run] [Cancel] [More]	
example_bash_operator	airflow	0	00:00:00		2023-06-19, 00:00:00	0	[Run] [Cancel] [More]	
example_branch_datetime_operator	airflow	0	@daily		2023-06-19, 00:00:00	0	[Run] [Cancel] [More]	
example_branch_datetime_operator_2	airflow	0	@daily		2023-06-19, 00:00:00	0	[Run] [Cancel] [More]	
example_branch_dop_operator_v3	airflow	0	00:00:00		2023-06-20, 18:36:00	0	[Run] [Cancel] [More]	
example_branch_labels	airflow	0	@daily		2023-06-19, 00:00:00	0	[Run] [Cancel] [More]	
example_branch_operator	airflow	0	@daily		2023-06-19, 00:00:00	0	[Run] [Cancel] [More]	
example_branch_python_operator_decorator	airflow	0	@daily		2023-06-19, 00:00:00	0	[Run] [Cancel] [More]	
example_complex	airflow	0	@daily			0	[Run] [Cancel] [More]	

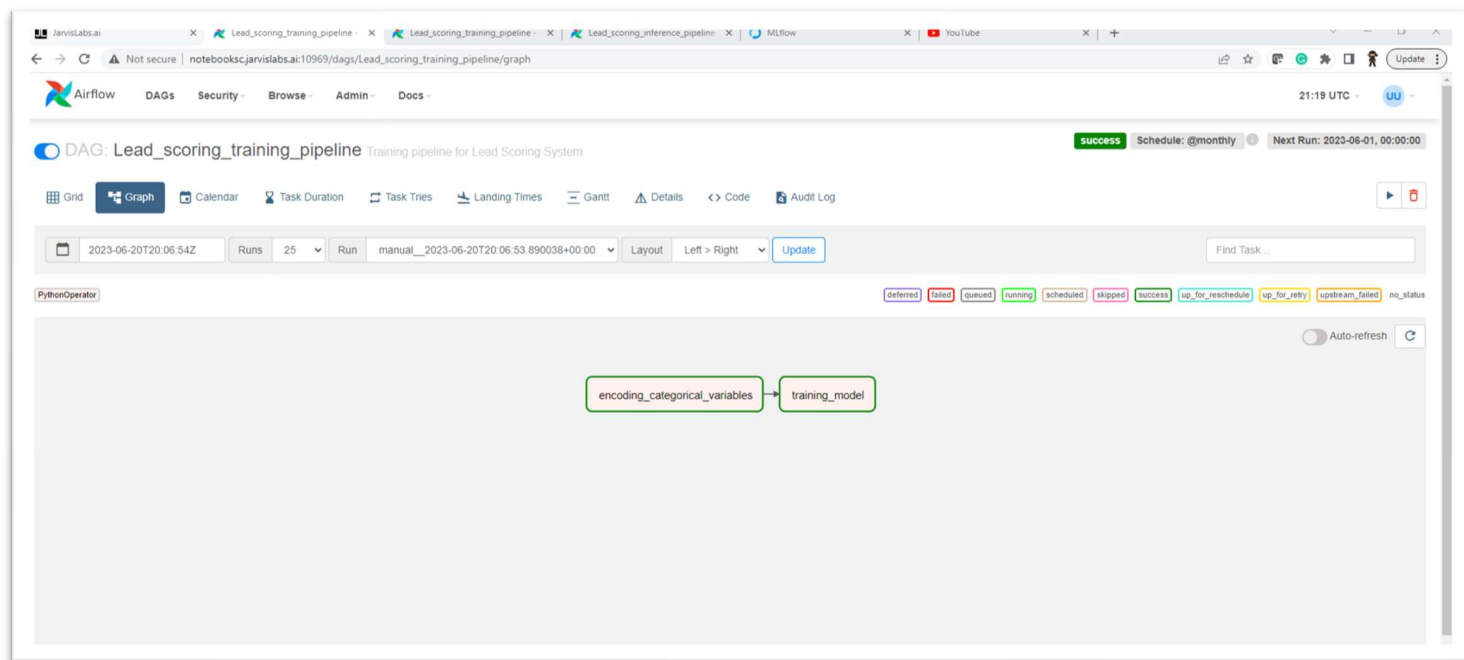
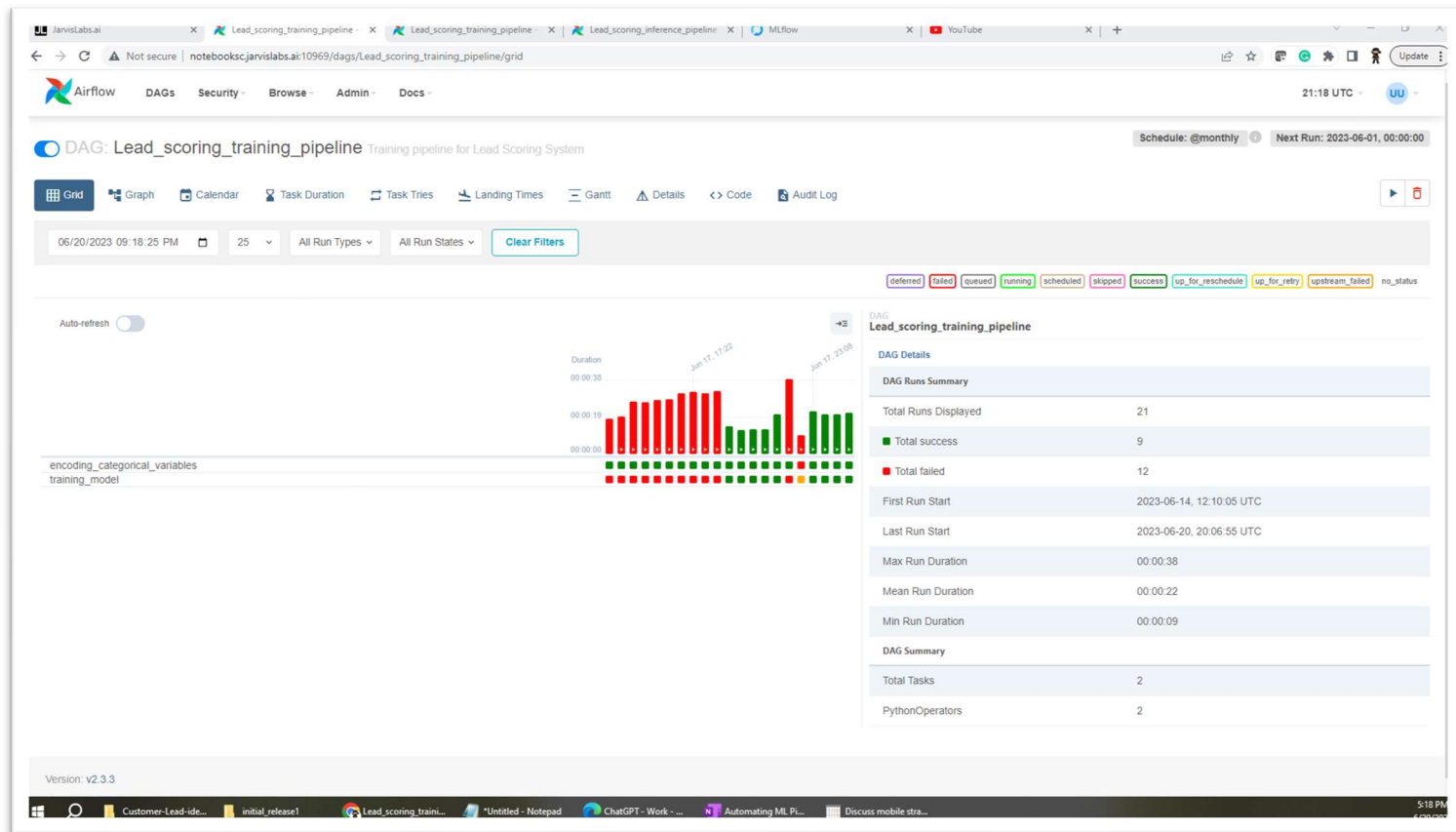
2. Lead_Scoring_Data_Engineering_Pipeline



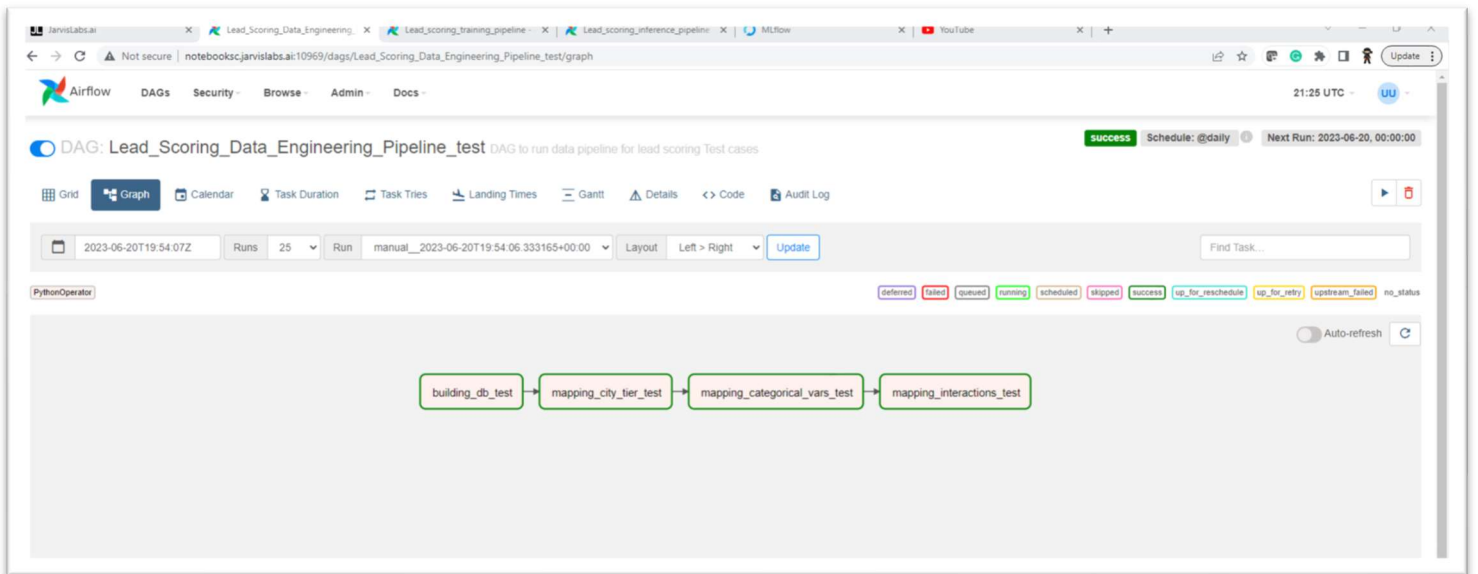
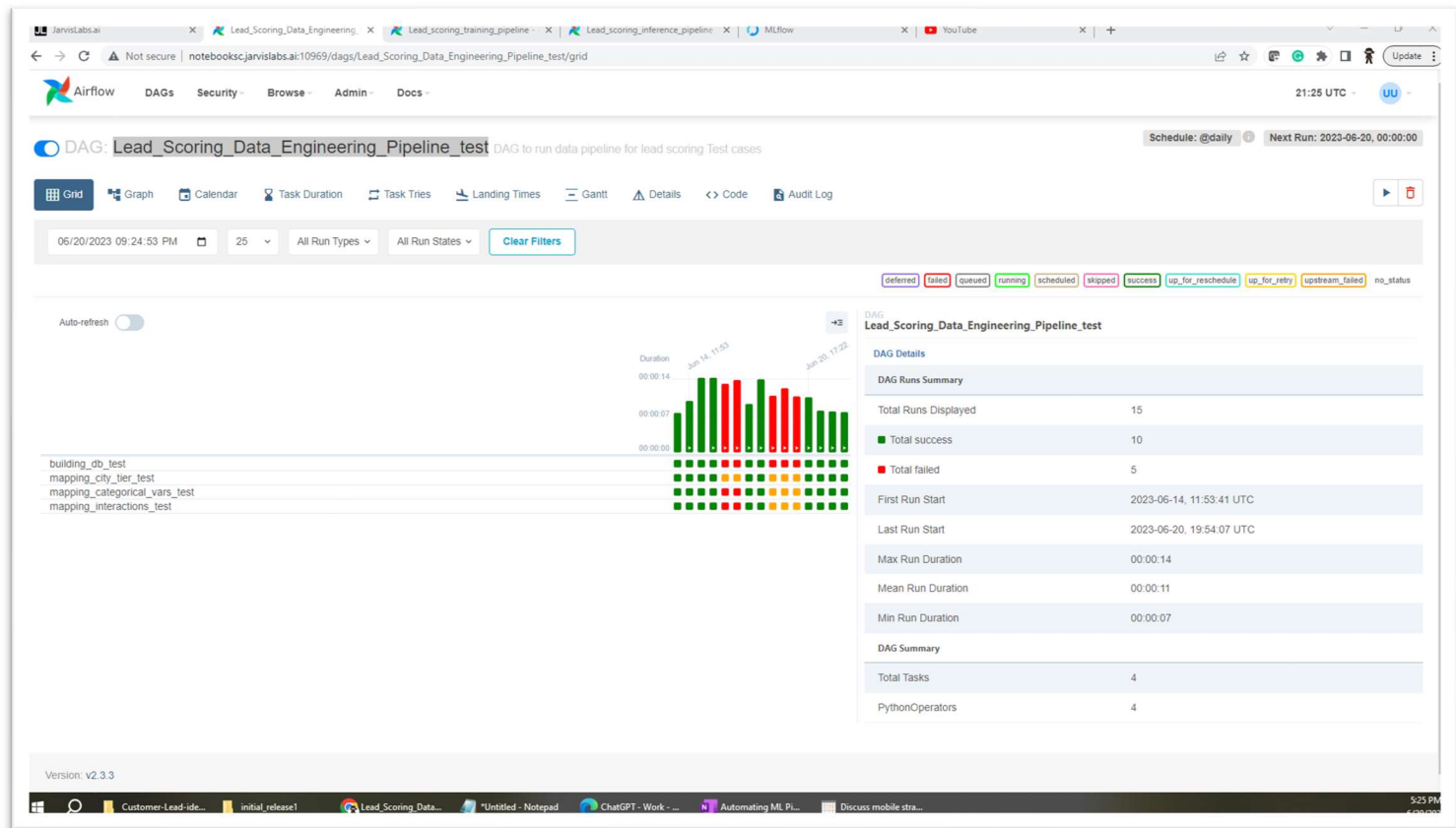
3. Lead_Scoring_Data_Engineering_Pipeline



4. Lead_Scoring_Data_Engineering_Pipeline



5. Lead_Scoring_Data_Engineering_Pipeline_test



6. mlFlow Lead Scoring Model Experimentation

The screenshot shows the mlflow Experiments page for the experiment 'Lead_scoring_model_experimentation_0616'. The page displays a list of 44 matching runs. The runs are sorted by start time, with the most recent runs at the top. The table columns include Start Time, Duration, Run Name, User, Source, Version, Models, Metrics (AUC, Accuracy, F1), Parameters (C, CPU Jobs, Categorical Feat), and Tags (Source). The runs are grouped by their start time, with runs from 53 minutes ago, 48 minutes ago, 49 minutes ago, and 3 days ago.

Start Time	Duration	Run Name	User	Source	Version	Models	AUC	Accuracy	F1	C	CPU Jobs	Categorical Feat	Source
53 minutes ago		Session Init...	root	ipykernel...	-	-	-	-	-	-	-1	4	setup
48 minutes ago		Light Gradi...	root	ipykernel...	-	sklearn	0.821	0.738	0.762	-	-	-	create_model
49 minutes ago		Naive Bayes	root	ipykernel...	-	sklearn	0.738	0.679	0.728	-	-	-	compare_m...
49 minutes ago		Ridge Classi...	root	ipykernel...	-	sklearn	0	0.715	0.742	-	-	-	compare_m...
49 minutes ago		Linear Discr...	root	ipykernel...	-	sklearn	0.79	0.715	0.742	-	-	-	compare_m...
49 minutes ago		Logistic Reg...	root	ipykernel...	-	sklearn	0.792	0.717	0.741	1.0	-	-	compare_m...
49 minutes ago		Decision Tre...	root	ipykernel...	-	sklearn	0.817	0.736	0.758	-	-	-	compare_m...
49 minutes ago		Extra Trees C...	root	ipykernel...	-	sklearn	0.818	0.737	0.758	-	-	-	compare_m...
49 minutes ago		Random For...	root	ipykernel...	-	sklearn	0.819	0.737	0.76	-	-	-	compare_m...
49 minutes ago		Light Gradi...	root	ipykernel...	-	sklearn	0.821	0.738	0.762	-	-	-	compare_m...
49 minutes ago		Extreme Gra...	root	ipykernel...	-	-	-	-	-	-	-	-	-
3 days ago		Session Init...	root	ipykernel...	-	-	-	-	-	-	-1	4	setup
3 days ago		Light Gradi...	root	ipykernel...	-	sklearn	0.821	0.738	0.762	-	-	-	create_model
3 days ago		Naive Bayes	root	ipykernel...	-	sklearn	0.738	0.679	0.728	-	-	-	compare_m...
3 days ago		Ridge Classi...	root	ipykernel...	-	sklearn	0	0.715	0.742	-	-	-	compare_m...
3 days ago		Linear Discr...	root	ipykernel...	-	sklearn	0.79	0.715	0.742	-	-	-	compare_m...
3 days ago		Logistic Reg...	root	ipykernel...	-	sklearn	0.792	0.717	0.741	1.0	-	-	compare_m...

The screenshot shows the mlflow Experiments page for the experiment 'Lead_scoring_model_experimentation_0616', specifically the details of a run named 'Light Gradient Boosting Machine'. The run is dated 2023-06-20 16:38:57 and has a status of UNFINISHED. The source is 'ipykernel_launcher.py' and the user is 'root'. The lifecycle stage is 'active' and the parent run is 'ab4fe1f7b60d48dcbe376b97b3989e2'.

Light Gradient Boosting Machine

Date: 2023-06-20 16:38:57
Status: UNFINISHED
Source: ipykernel_launcher.py
User: root
Lifecycle Stage: active
Parent Run: ab4fe1f7b60d48dcbe376b97b3989e2

Metrics (8)

Name	Value
AUC	0.821
Accuracy	0.738
F1	0.762
Kappa	0.476
MCC	0.485
Prec.	0.702
Recall	0.833
TT	1.89

Tags (5)

Artifacts

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
subsample_freq	0

Kappa0.476

MCC0.485

Prec0.702

Recall0.833

TT1.89

Tags (5)

Artifacts

model

MLModel

conda.yaml

model.pkl

python_env.yaml

requirements.txt

AUC.png

Confusion Matrix.png

Feature Importance.png

Holdout.html

Full Path:/home/mlruns/5/849e4d8af02c473899228d29ec7055df/artifacts/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:/849e4d8af02c473899228d29ec7055df/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()
```

Predict on a Pandas DataFrame

```
import mlflow
logged_model = 'runs:/849e4d8af02c473899228d29ec7055df/model'

# 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))
```

Customer-Lead-ide...

initial_release1

MLflow - Google C...

*Untitled - Notepad

ChatGPT - Work - ...

Automating ML Pl...

Discuss mobile stra...

5:29 PM

JarvisLabs.ai

Lead_Scoring_Data_Engineering

Lead_scoring_training_pipeline

Lead_scoring_inference_pipeline

MLflow

YouTube

Not secure | notebooksc.jarvislabs.ai:10968/#/models

mlflow 1.26.1

Experiments

Models

GitHub

Docs

Registered Models

Share and manage machine learning models. Learn more

Create Model

Search by model name

Search

Filter

Clear

Name	Latest Version	Staging	Production	Last Modified	Tags
LightGBM	Version 28	-	Version 1	2023-06-20 16:07:15	-
lightgbm	Version 1	-	Version 1	2023-06-15 00:11:02	-

< 1 > 10 / page

JarvisLabs.ai

Lead_Scoring_Data_Engineering

Lead_scoring_training_pipeline

Lead_scoring_inference_pipeline

MLflow

YouTube

Not secure | notebooksc.jarvislabs.ai:10968/#/models/LightGBM

mlflow 1.26.1

Experiments

Models

GitHub

Docs

Registered Models > LightGBM

LightGBM

Created Time: 2023-06-15 00:09:02

Last Modified: 2023-06-20 16:07:15

Description Edit

Tags

Versions All Active 1 Compare

Version	Registered at	Created by	Stage	Description
<input type="checkbox"/> Version 28	2023-06-20 16:07:15		None	
<input type="checkbox"/> Version 27	2023-06-20 14:18:08		None	
<input type="checkbox"/> Version 26	2023-06-20 14:15:57		None	
<input type="checkbox"/> Version 25	2023-06-17 19:08:48		None	
<input type="checkbox"/> Version 24	2023-06-17 19:03:08		None	
<input type="checkbox"/> Version 23	2023-06-17 19:02:51		None	
<input type="checkbox"/> Version 22	2023-06-17 18:59:49		None	
<input type="checkbox"/> Version 21	2023-06-17 17:08:38		None	
<input type="checkbox"/> Version 20	2023-06-17 17:04:13		None	
<input type="checkbox"/> Version 19	2023-06-17 16:51:15		None	

< 1 2 3 >

notebooksc.jarvislabs.ai:10968/#/

Customer-Lead ide... initial release1 MLflow - Google C... *Untitled - Notepad ChatGPT - Work - ... Automating ML Pl... Discuss mobile stra... 5:32 PM

mlflow 1.26.1 Experiments Models GitHub Docs

Experiments

Search Experiments

- Default
- Model_Building_Pip...
- Model_Building_Pip...
- Old_Lead_scoring_m...
- Lead_scoring**
- Lead_scoring_model...

Experiment ID: 4

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 83 matching runs

	Start Time	Duration	Run Name	User	Source	Version	Models	AUC	Accuracy	F1	C	CPU Jobs	Categorical Feat	Source	URI	URI
	1 hour ago	50s	Session Init...	root	ipykernel...	-		-	-	-	-	-1	5	setup	06abc900	9cc7
	1 hour ago	50s	run_LightGB	root	airflow	-	LightGBM/28	-	-	-	-	-	-	-	-	-
	3 hours ago	42s	run_LightGB	root	airflow	-	LightGBM/27	-	-	-	-	-	-	-	-	-
	3 hours ago	42s	run_LightGB	root	airflow	-	LightGBM/26	-	-	-	-	-	-	-	-	-
	2 days ago	50s	run_LightGB	root	airflow	-	LightGBM/25	-	-	-	-	-	-	-	-	-
	2 days ago	48s	run_LightGB	root	airflow	-	LightGBM/24	-	-	-	-	-	-	-	-	-
	2 days ago	45s	run_LightGB	root	airflow	-	LightGBM/23	-	-	-	-	-	-	-	-	-
	2 days ago	44s	run_LightGB	root	airflow	-	LightGBM/22	-	-	-	-	-	-	-	-	-
	3 days ago	36s	run_LightGB	root	airflow	-	LightGBM/21	-	-	-	-	-	-	-	-	-
	3 days ago	39s	run_LightGB	root	airflow	-	LightGBM/20	-	-	-	-	-	-	-	-	-
	3 days ago	37s	run_LightGB	root	airflow	-	LightGBM/19	-	-	-	-	-	-	-	-	-
	3 days ago	40s	run_LightGB	root	airflow	-	LightGBM/18	-	-	-	-	-	-	-	-	-
	3 days ago	51s	run_LightGB	root	airflow	-	LightGBM/17	-	-	-	-	-	-	-	-	-
	3 days ago	52s	run_LightGB	root	airflow	-	LightGBM/16	-	-	-	-	-	-	-	-	-
	3 days ago	52s	run_LightGB	root	airflow	-	LightGBM/15	-	-	-	-	-	-	-	-	-
	3 days ago	52s	run_LightGB	root	airflow	-	LightGBM/14	-	-	-	-	-	-	-	-	-
	3 days ago	50s	run_LightGB	root	airflow	-	LightGBM/13	-	-	-	-	-	-	-	-	-
	3 days ago	51s	run_LightGB	root	airflow	-	LightGBM/12	-	-	-	-	-	-	-	-	-
	3 days ago	51s	run_LightGB	root	airflow	-	LightGBM/11	-	-	-	-	-	-	-	-	-
	3 days ago	51s	run_LightGB	root	airflow	-	LightGBM/10	-	-	-	-	-	-	-	-	-
	3 days ago	39s	Session Init...	root	ipykernel...	-		-	-	-	-	-1	5	setup	00b0321a	fa03
	3 days ago	39s	run_LightGB	root	airflow	-	LightGBM/9	-	-	-	-	-	-	-	-	-
	3 days ago	39s	run_LightGB	root	airflow	-	LightGBM/8	-	-	-	-	-	-	-	-	-

run_LightGB

Date: 2023-06-20 16:07:10 Source: airflow User: root

Duration: 50s Status: FINISHED Lifecycle Stage: active

Description Edit

Parameters (20)

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

jarvislabs.ai

Lead_Scoring_Data_Engineering

Lead_scoring_training_pipeline

Lead_scoring_training_pipeline

MLflow

YouTube

Not secure | notebooks.jarvislabs.ai:10968/#/experiments/4/runs/bf259cc9a4f54928b05097ce77659566

client: warn

subsample: 1.0

subsample_for_bin: 200000

subsample_freq: 0

Metrics (1)

Name	Value
auc_bst	0.739

Tags

Artifacts

models

MLmodel

conda.yaml

model.pkl

python_env.yaml

requirements.txt

Full Path:/home/miniruns/4/bf259cc9a4f54928b05097ce77659566/artifacts/models

LightsGBM v28 Registered on 2023-06-20

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:/bf259cc9a4f54928b05097ce77659566/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:/bf259cc9a4f54928b05097ce77659566/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))
```