IIn this segment you will create a pipeline to automate inference using Airflow

For creating the pipeline follow the below mentioned instructions:

1. Go to the **Assignment/03\_inference\_pipeline**/scripts folder.

2. First we will break down all the **model inference tasks into functions in the utils.py file** you have been provided with functions level breakdown of all the tasks. We will basically need 2

Functions for inference. The first function will encode the features and the second function will generate models prediction on the data and write it to the db. On top of this we will also do some sanity checks.

You can **copy the encode\_data function** that you used in the training pipeline but there will be some **modification**. Recall that in the encode\_data function that you used, you had to save the output columns as a list. In case of training this includes “app\_complete\_flag” which is available to us during training but will not have that inference so **you modify the encode\_data function accordingly.**

3.Once you have written the code into the respective functions. You will **modify the function to read their input data and write their output to the database present in file 'lead\_scoring\_data\_cleaning.db'**. The function will be executed in the following order so make sure to read the input for the functions appropriately

**encode\_data -> input\_features\_check -> load\_model -> prediction\_col\_check**

**Ensure that you are reading and writing from the appropriate tables in the db.**

NOTE: We need to clean our inference data before we can use it to get models inference. Thus we will run the data pipeline on the raw inference data and use the cleaned data for inference. Thus our **encode\_feature will read the input from data pipeline’s output which is present in the ‘lead\_scoring\_data\_cleaning.db’ file**. Also note that since we are not creating any db here we do not require build\_db function

Once you have coded the functions, create a notebook and import utils file in it and run all the functions in the proper order. This will help you **debug your code**

NOTE : Before you run the get\_model\_prediction function ensure that you have run mlflow server command from the terminal at the appropriate path. Also ensure that **you have promoted the registered model to promotion before running get\_models\_prediction function** as this function loads model which is in production stage(“Lead\_scoring\_mlflow\_production”)

5.Once you have ensured that the functions in utils.py file are working as intended, **store all the constant values in constants.py** and repeat step 3 to ensure that the code is working properly

6. Now **create a pipeline** using airflow in the 'lead\_scoring\_inference\_pipeline.py'. Follow the instructions present in the file for creating the dag for airflow.

7.Now go to the '~/airflow/dags/' folder and create a folder named 'Lead\_scoring\_inference\_pipeline' in it and copy paste the following files in it.The following file should be present in **Lead\_scoring\_inference\_pipeline folder**

├── lead\_scoring\_inference\_pipeline.py

├── constants.py

└── utils.py

8. **After copying all the necessary files and folder from 'scripts' folder to 'Lead\_scoring\_inference\_pipeline' and modify your constants, utils and lead\_scoring\_inference\_pipeline.py as the paths have been changed when you have changed the directory. Make changes in the file path names in constants.py file.**

9. With this you should be able to run the inference pipeline. Follow the instructions mentioned in the Airflow module to run the inference pipeline and **take the screenshot of Airflow UI for submission.**

8.**Manually trigger the inference pipeline and see if we are getting the expected results.**

9.Now **copy the “leadscoring\_inference.csv” file** present in “Assignment/03\_inference\_pipeline/Data” to the “airflow/dags/Lead\_scoring\_data\_pipeline/data” folder and modify the constants.py in the Lead\_scoring\_data\_pipeline such that the data pipeline read “leadscoring\_inference.csv” instead of “leadscoring.csv”.

NOTE: “leadscoring\_inference.csv” is our inference dataset and it does not contain “app\_complete\_flag”.

NOTE: We need to clean our inference data before we can use it to get models inference. Thus we will run the data pipeline on the raw inference data and use the cleaned data for inference.

10. Once this is done first trigger the Data pipeline manually and then trigger the inference pipeline once the data pipeline’s run is complete. **Revise till you get the expected output. Follow the instructions mentioned in the Airflow module to run the Inference pipeline and take the screenshot of Airflow UI for submission.** 10.Pat yourself on the back as you have completed the assignment.

!!!WOOOOHOOOOOO YOU HAVE COMPLETED THIS ASSIGNMENT!!!

!!!!!!!!!!!!!!!!!!WELL DONE!!!!!!!!!!!!!!!!!!!!!!