W2 - Data Ops, MLOps, AIOps

Graded Assignment

Objective:

To use mlflow for logging and querying machine learning experiments. The task is to build, log, and track multiple versions of a classification model that aims to predict the passengers who survived the titanic shipwreck.

Context:

Developing a machine learning model is an iterative process. It involves various rounds of training and evaluation. There are multiple tools that are used for streamlining the machine learning operations. We will be using mlflow for logging and tracking the experiments which will help us to analyze how accurate the predictions are being made and also compare the performance of the multiple experiments and we can tweak the model parameters to ensure the best performance.

Dataset Description:

We will be using the titanic dataset.

- The sinking of the Titanic is one of the most infamous shipwrecks in history.
- On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.
- While there was some element of luck involved in surviving, it seems some

groups of people were more likely to survive than others.

- Please refer to the Dataset Description.pdf document for more detailed information of the variables.

Approach:

- We have data of titanic passengers across attributes like 'PassengerId', 'Name', 'Sex', 'Age', ticket class, fare, whether the passenger survived or not(target) etc.
- There are redundant columns that need to be dropped.
- There are missing values that need to be taken care of.
- There will be a few categorical variables that we will have to encode.
- Using this data, we will attempt to train a classification model that predicts which passengers survived the titanic shipwreck.
- We will evaluate the model performance using scores of accuracy, precision, recall and f1 score.
- We will log parameters, metrics, models, and run different versions of the machine learning experiments using mlflow package and track and compare the performance using the mlflow ui.

Steps (50 points):

- Install mlflow and required libraries. (5 points)
- Load the given data, perform basic EDA, and build a classification model to predict the passengers who survived the titanic shipwreck. (Note:- you can use any classification algorithm you have learned in your previous modules). (20 points)
- Log the following parameters using mlflow. (5 points)

a. Hyperparameters (i.e. no. of estimators in case of random forest) if any
b. Evaluation metrics like accuracy, precision, recall, f1 score.

- Run the above model and track the logged parameters in the mlflow ui. (5 points)
- Make some changes in the model training (like- change a few hyperparameters, train size, etc) and train the updated model. (10 points)
- Track the logged parameters of the second version of the model and compare it with the first model. (5 points)
- Serve the model using ML model format with mlflow to deploy a local REST server and pass some sample data and see the predictions. (optional)

Submission:

- Your submission should include a document having screenshots of all the steps and the .py (or ipynb or .html) file of the model.