Tên: Bùi Vĩnh Tiến

Lớp: MSEHCM20

Môn: DDM501c

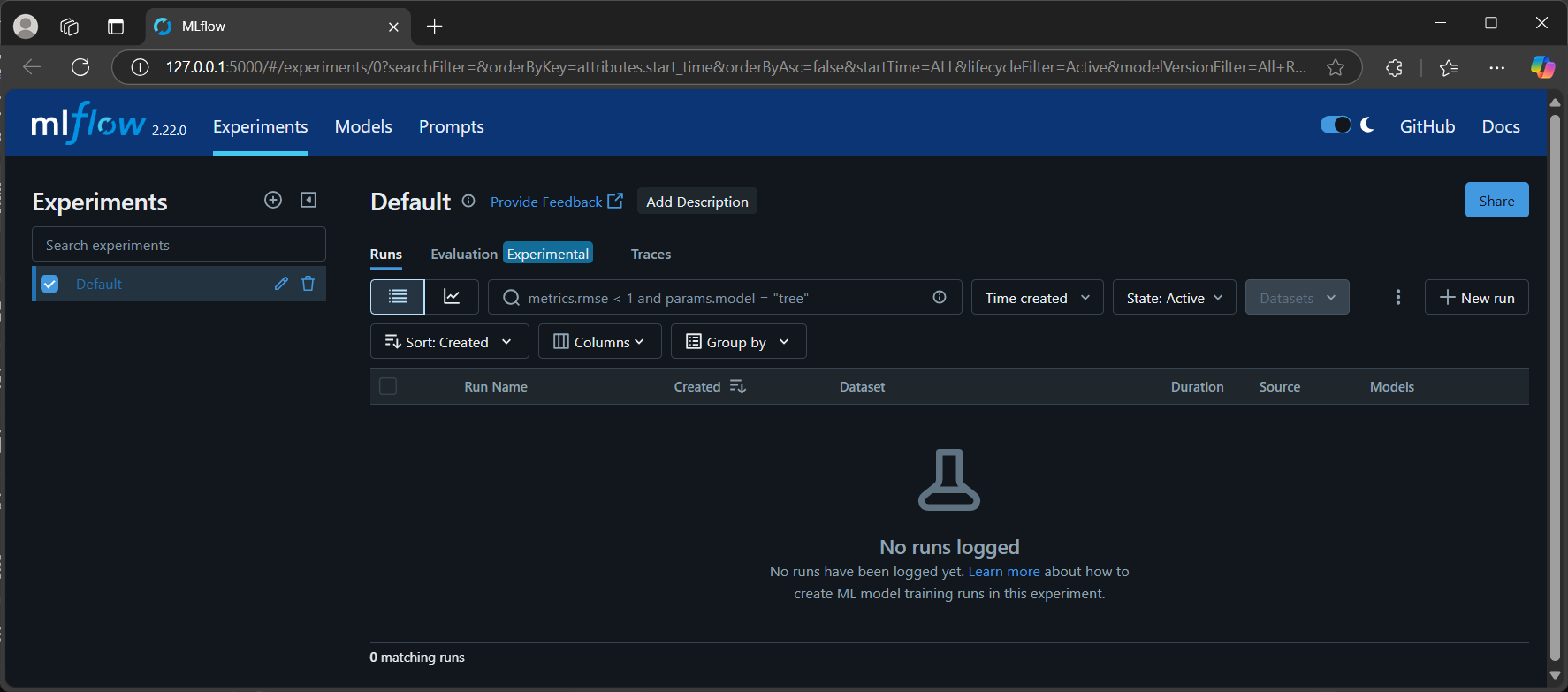
Các bước thực hiện dự án:

1. Tham khảo lại luồng xử lí tổng quan: Data -> Dev -> ML Ops
2. Tạo tệp chứa toàn bộ các thư viện: requirements.txt, khởi tạo môi trường ảo.
3. Tạo API sinh dữ liệu.

* Tạo chương trình tạo ra dữ liệu dataGen bằng make\_classification.
* Dữ liệu không hợp lý, số lẻ quá nhiều, cần điều chỉnh để tạo ra dữ liệu có ý nghĩa hơn
* Lưu data thành một bộ (data/loan\_prediction\_data.sv để so sánh giữa các mô hình trên cùng một bộ dữ liệu

1. Khởi tạo một mlflow trên mạng local từ Terminal: mlflow-server

Hệ thống báo rỗng khi chưa có lượt huấn luyện dữ liệu nào



1. Huấn luyện:

* Xử lý data -> các vị trí dữ liệu không phải là kiểu số (quê quán, giới tính,…)
* Đọc file data, chia ra 2 tập train và test
* Huấn luyện trên tập train sử dụng:
  + Logistic Regression (nhiều tham số C, solver)
  + GridSearchCV (Logistic Regression tìm tham số tốt nhất)
  + Random Forest Classifier

Xử lí lỗi

* Xử lí lỗi khi huấn luyện với dữ liệu dạng string (tên tỉnh thành, giới tính, trạng thái hôn nhân).
* Xử lí lỗi file artifact của mlflow -> sai đường dẫn-> cập nhật lại

1. Tạo API quét và tìm model tốt nhất trong các model đã huấn luyện dựa trên accuracy.
2. Tạo ứng dụng để sử dụng mô hình đã huấn luyện app.py.

* Lỗi không kích hoạt được Server. SimpleFlask chạy => lỗi có thể đến từ phần load model từ mlflow -> xác nhận lỗi này: khi chuyển qua tải thẳng model thông qua đường dẫn file thì chương trình chạy được.
* Cần xem lại thiết lập mlflow
* Khởi chạy mlflow Server thay cho chỉ mlflow UI -> cập nhật lại bước 4

1. Tạo lại Makefile để thuận tiện sử dụng.
2. Đẩy lên Git