ConQ999

Project: MLDockFlow
Upgraded Project Presentation

ConQ999 Team

Thành viên Team

ConQ999

Module 5 nhóm có 6 thành viên chính thức

Đàm Nguyên Khánh	Leader
Vũ Thái Sơn	Tech leader
Phạm Khánh Quân	Member
Võ Hoàng	Member
Nguyễn Quang Linh	Member

Quản lý team: **Discort**

Các công cụ sử dụng cho AIO Conquer: **Overleaf** | **GG Colab**

Table of contents

ConQ999

ConQ999

01

Objectives of the project

ConQ999

Vấn đề hiện tại

Data scientists thường gặp các khó khăn sau:

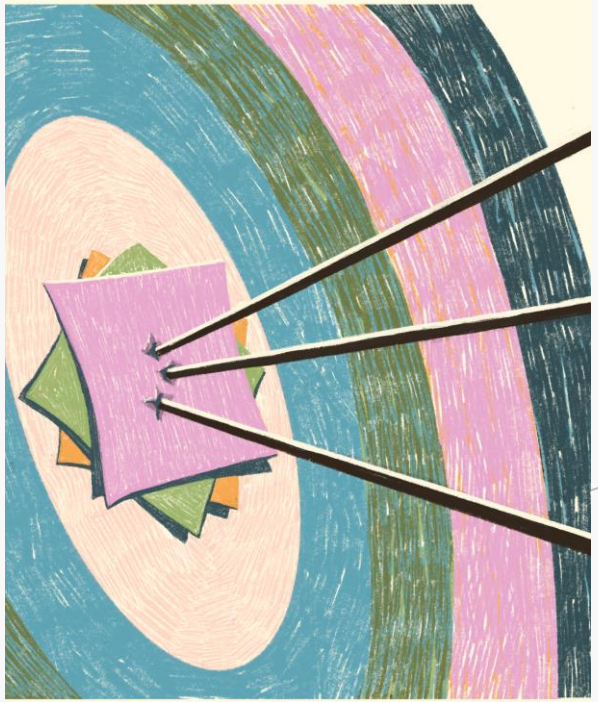
1. Thời gian lớn dành cho preprocessing & tuning.
2. Thiếu hệ thống theo dõi thí nghiệm có tổ chức.
3. Khó tái lập và giám sát pipeline huấn luyện.
4. Chưa có giải pháp triển khai thống nhất.

ConQ999

Mục tiêu MLDockFlow

All-in-One MLOps Solution:

1. Tự động hóa toàn bộ pipeline từ preprocessing → training → tracking → deployment.
2. Tích hợp MLflow để theo dõi và quản lý thí nghiệm.Hỗ trợ nhiều mô hình (Linear, Ridge, XGBoost, CatBoost, LGBM).
3. Tối ưu siêu tham số với Optuna.Container hóa và triển khai qua Docker Compose.



Objectives

ConQ999

Our aim

MLDockFlow hướng tới việc chuyển một dự án học thuật thành hệ thống MLOps hoàn chỉnh.

The goal

- Hệ thống học máy có thể tái lập, giám sát và mở rộng.
- Dễ bảo trì, minh bạch trong quản lý thí nghiệm.
- Giao diện thân thiện cho người dùng phi kỹ thuật.

ConQ999

02

Structure of the project

Cấu trúc Dự án MLDockFlow

ConQ999

```
AIO2025_Project5.1_HousesPricing-main/
│
├── 📁 data/                          # Dữ liệu
│   └── raw/                          # Dữ liệu gốc (không thay đổi)
│       └── train-house-prices-advanced-regression-techniques.csv
│
├── 📁 src/                           # Source code chính
│   │
│   ├── 📁 api/                       # FastAPI Application
│   │   ├── __init__.py
│   │   ├── main.py                   # FastAPI app, endpoints định nghĩa
│   │   ├── models.py                 # Pydantic models cho request/response
│   │   ├── inference.py              # Inference logic & CLI tool
│   │   ├── run_api.py                # Script khởi chạy API server
│   │   ├── test_api.py               # Test script cho API endpoints
│   │   ├── Dockerfile                # Docker image cho API service
│   │   └── README.md                 # Tài liệu API
│   │
│   ├── 📁 processing/                # Xử lý dữ liệu
│   │   ├── __init__.py
│   │   ├── transformers.py           # Custom transformers (OrdinalMapper,
MissingnessIndicator, etc.)
│   │   └── data_processing.py        # Preprocessing pipeline
│   │
│   ├── 📁 e_featuring/               # Feature Engineering
│   │   ├── __init__.py
│   │   └── data_featuring.py         # Domain-specific features (18 features)
│   │
```

Project structure

Modular hóa theo 4 tầng: processing – training – monitoring – deployment giúp quản lý toàn bộ vòng đời mô hình từ tiền xử lý dữ liệu, huấn luyện, theo dõi bằng MLflow, đến triển khai qua Docker Compose với FastAPI và Streamlit UI.

Cấu trúc Dự án Email Classifier

ConQ999

```
├── training/                    # Training Model
│   ├── __init__.py
│   ├── pipeline.py              # Training pipeline builder
│   └── train_model.py           # Main training script với MLflow
│
├── configs/                     # Cấu hình
│   └── best_model_config.json   # Cấu hình model tốt nhất (hyperparameters, performance metrics)
│
├── frontend/                    # Streamlit UI
│   ├── app.py                   # Streamlit application
│   ├── Dockerfile               # Docker image cho frontend
│   ├── FRONTEND_DESIGN.md       # Thiết kế UI
│   └── README.md                # Tài liệu frontend
│
└── models/                      # Models đã train (auto-generated, không commit)
    ├── best_pipeline.joblib     # Pipeline hoàn chỉnh (features + model)
    └── feature_pipeline.joblib  # Feature engineering pipeline

├── deployments/                 # Deployment configurations
│
│   ├── api/                     # API deployment
│   │   └── docker-compose.yaml  # Docker Compose cho API + Frontend + MLflow
│   │   └── Dockerfile           # (tùy chọn, có thể dùng từ src/api/)
│   │
│   └── mlflow/                  # MLflow tracking server
│       └── docker-compose.yaml  # MLflow standalone server
│
├── notebooks/                   # Jupyter Notebooks
│   ├── house_price_analysis.ipynb      # Exploratory data analysis
│   └── house_price_analysis_mlflow.ipynb  # Experiments với MLflow tracking
│
├── train.py                     # Script training chính (entry point)
├── requirements.txt             # Python dependencies
├── README.md                    # Tài liệu chính của dự án
├── .gitignore                   # Git ignore rules
├── .gitattributes               # Git attributes
└── .pre-commit-config.yaml      # Pre-commit hooks
```

Core Components

ConQ999

Data Pipeline

Custom transformers, feature engineering, scaling

Training Pipeline

Tự động huấn luyện & log metrics bằng MLflow

Model Registry

Quản lý version & reproducibility

API Service

FastAPI REST endpoints

UI Layer

Streamlit web interface

ConQ999

03

Upgrade of the project

Các nâng cấp

ConQ999

MLflow Integration

Theo dõi parameters, metrics, artifacts.

Optuna Optimization

Tối ưu siêu tham số tự động.

Stacking Ensemble

Kết hợp nhiều mô hình tăng hiệu năng.

Docker Compose Deployment

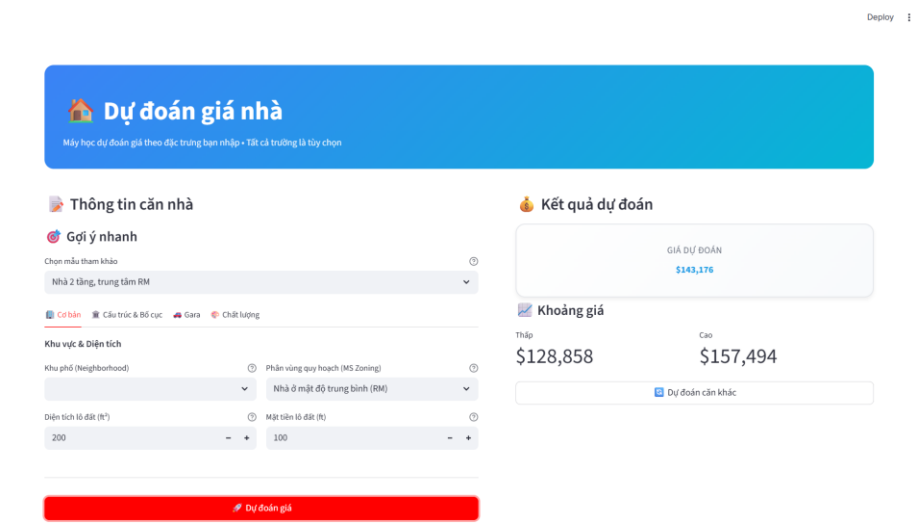
Tạo môi trường thống nhất.

Streamlit UI

Giao diện web tương tác trực quan.

ConQ999

Streamlit UI



Trải nghiệm người dùng

- Input form nhập thông tin nhà
- Preset mẫu: Nhà nhỏ / trung bình / cao cấpReal-time prediction & kết quả trực quan
- Responsive design
- Kết nối trực tiếp FastAPI qua API_URL

Không cần chạy notebook – chỉ mở trình duyệt để dự đoán.

ConQ999

04

Result

ConQ999

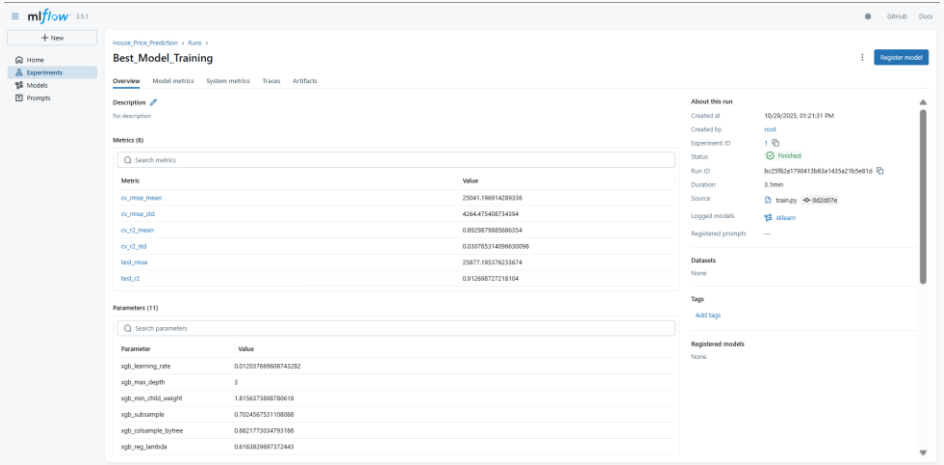
Kết quả Training

Model	Test RMSE	Test R^2
Single models		
Ridge (baseline)	27.763,37	0,889.969
Lasso (baseline)	34.402,87	0,831.048
LinearRegression (baseline)	55.032,41	0,567.676
XGB	**24.608,889.79**	**0,921.046.714**
CatBoost	27.138,542.01	0,903.980.555
LGBM	28.937,452.11	0,890.829.137
RandomForest	29.694,111.37	0,885.045.274
Ridge	31.564,712.94	0,870.105.772
ElasticNet	31.820,724.43	0,867.990.164
Lasso	33.069,627.43	0,857.424.544
SVR	88.551,152.4	−0,022.291.149
Ensembles (Stacking)		
CatBoost+RandomForest+Ridge	**27.682,867.28**	**0,900.090.15**
CatBoost+LGBM+Ridge	27.966,034.36	0,898.035.748
CatBoost+RandomForest+LGBM	28.372,076.97	0,895.053.388
CatBoost+XGB+LGBM	28.305,464.5	0,895.545.6
CatBoost+XGB+RandomForest	28.496,588.92	0,894.130.242
CatBoost+XGB+Ridge	28.148,281.75	0,896.702.468
RandomForest+LGBM+Ridge	28.258,559.53	0,895.891.496
XGB+LGBM+Ridge	28.283,320.04	0,895.708.974
XGB+RandomForest+LGBM	28.501,769.42	0,894.091.746
XGB+RandomForest+Ridge	28.266,406.07	0,895.833.672

- XGBoost đạt hiệu năng cao nhất với **RMSE = 24,608 và R^2 = 0.921.**
- Các ensemble (Stacking) đem lại hiệu năng ổn định và khả năng tổng quát hóa tốt, **RMSE ≈ 27 k và R^2 ≈ 0.90**.
- So với phiên bản gốc, hiệu suất được cải thiện ~7% và đáp ứng chuẩn MLOps sẵn sàng triển khai.

ConQ999

Experiment Tracking



MLflow Tracking & Logging

- Log parameters, metrics, artifacts tự động.
- UI tại http://localhost:5555.
- Mỗi run gồm:
  - Parameters (learning_rate, n_estimators, ...).
  - Metrics (CV RMSE, Test RMSE, R^2).
  - Artifacts (model.pkl, pipeline.joblib).

MLflow giúp tái lập và so sánh các phiên bản mô hình trực quan.

Monitoring & Deployment

ConQ999

MLflow Tracking Server

Theo dõi thí nghiệm

FastAPI Service

Cung cấp REST API dự đoán

Streamlit UI

Giao diện demo người dùng

Healthcheck

Tự động giám sát API, restart container khi lỗi.

ConQ999

05

LIVE DEMO FEATURES

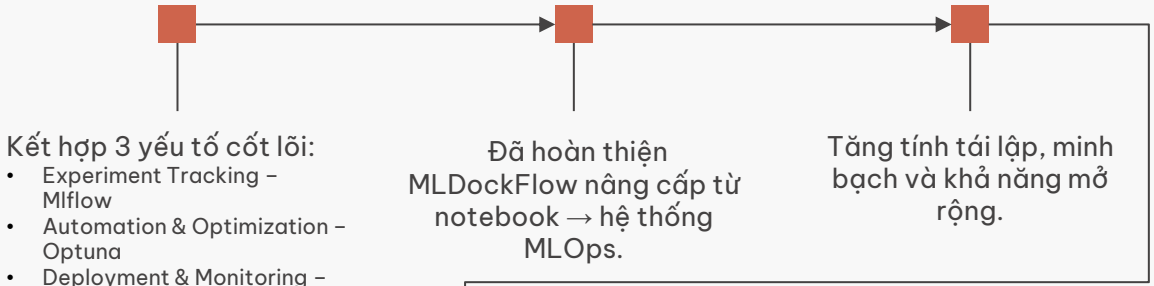
ConQ999

06

Conlusion

Kết quả & Hướng phát triển

ConQ999



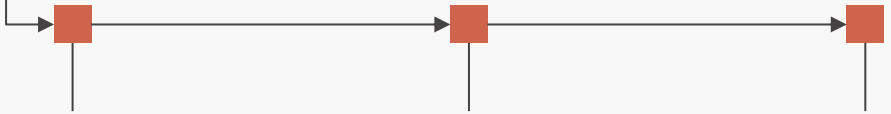
Kết hợp 3 yếu tố cốt lõi:

- Experiment Tracking – Mlflow
- Automation & Optimization – Optuna
- Deployment & Monitoring – Docker + FastAPI + Streamlit

Đã hoàn thiện MLDockFlow nâng cấp từ notebook → hệ thống MLOps.

Tăng tính tái lập, minh bạch và khả năng mở rộng.

HƯỚNG PHÁT TRIỂN



Mở rộng pipeline sang classification / time-series.

Giám sát bằng **Prometheus + Grafana.** Hỗ trợ **Kubernetes & Helm Chart** để scale.

Tự động retrain khi phát hiện data drift.

ConQ999

Thanks!

Do you have any questions?