



ĐẠI HỌC QUỐC GIA TP HCM
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

DỰ ĐOÁN XU HƯỚNG GIÁ CỔ PHIẾU GOOGLE DỰA TRÊN DỮ LIỆU GIÁ VÀ PHÂN TÍCH CẢM XÚC TIN TỨC BẰNG MACHINE LEARNING

PHAN GIA BẢO
MSSV: 2310251

GVHD: NGUYỄN AN KHƯƠNG

13/12/2025

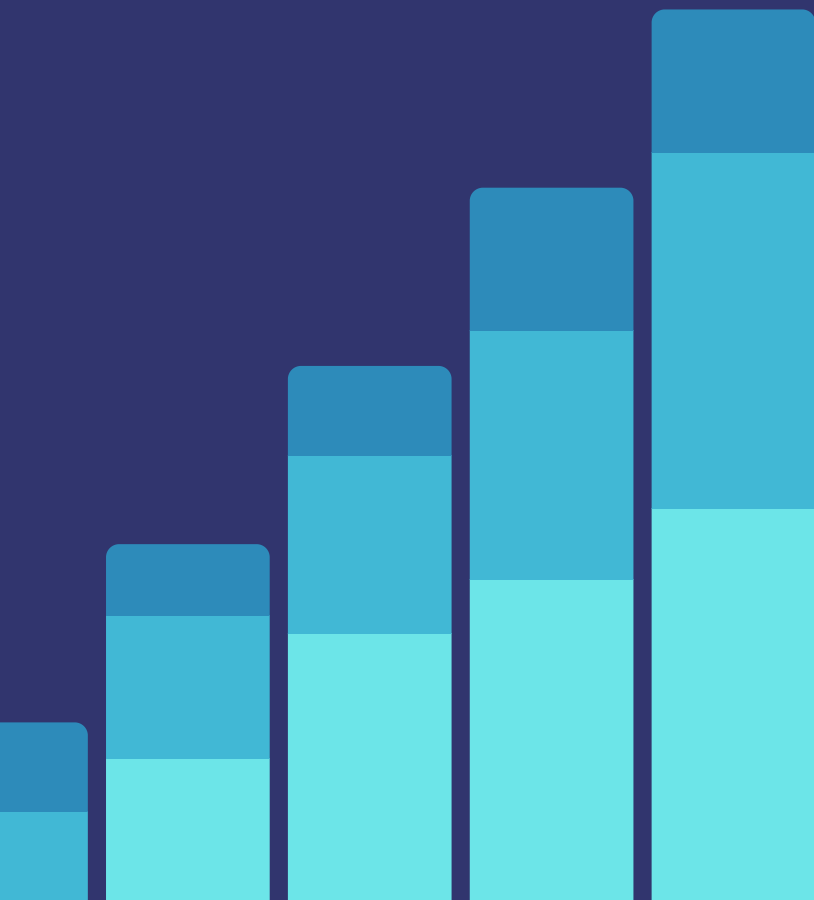
BỐI CẢNH



- Thị trường chứng khoán chịu ảnh hưởng đồng thời từ dữ liệu giá lịch sử và thông tin tin tức / cảm xúc nhà đầu tư.
- Các mô hình truyền thống chủ yếu dựa trên dữ liệu giá (technical indicators), trong khi nhiều nghiên cứu gần đây cho thấy sentiment từ tin tức và mạng xã hội có tác động đáng kể đến biến động giá cổ phiếu.

→ Đồ án này được thực hiện nhằm xây dựng và phân tích một pipeline dự đoán xu hướng giá cổ phiếu, từ đó đánh giá mức độ đóng góp của sentiment trong bài toán dự đoán tài chính.

MỤC TIÊU VÀ PHẠM VI BÀI TOÁN



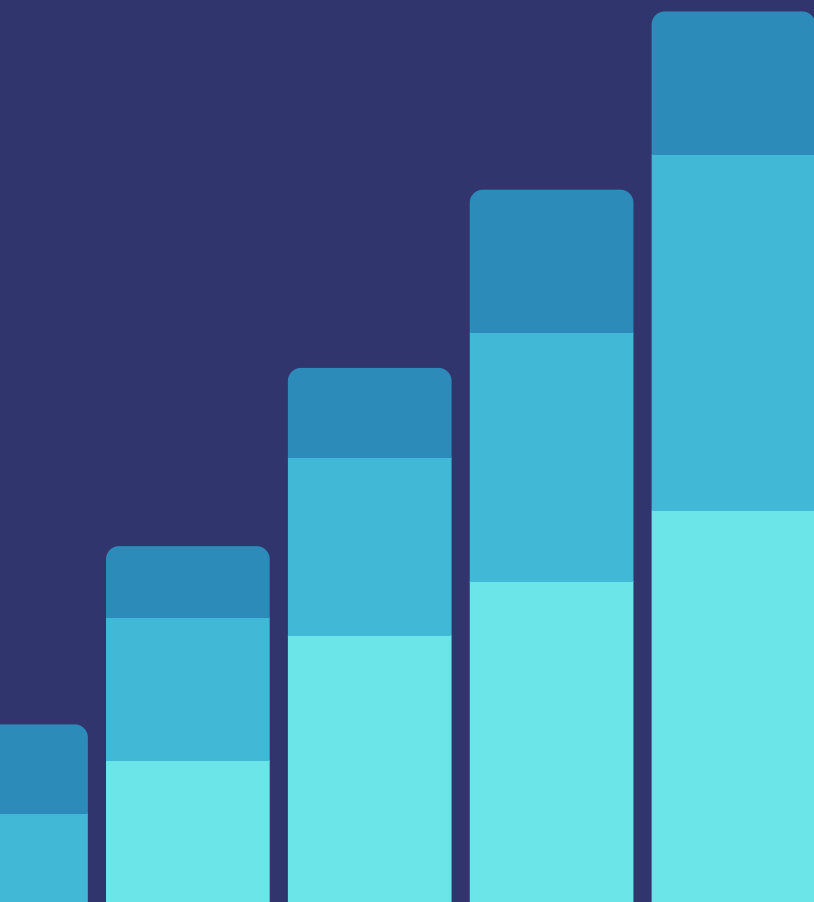
Mục tiêu

- Xây dựng pipeline Machine Learning end-to-end dự đoán xu hướng giá cổ phiếu Google (GOOGL) theo ngày
- Kết hợp dữ liệu giá cổ phiếu và dữ liệu tin tức thông qua phân tích cảm xúc (sentiment analysis)
- Đánh giá hiệu quả dự đoán của mô hình XG Boost trên dữ liệu thực tế
- Phân tích vai trò của sentiment tin tức trong việc cải thiện dự đoán xu hướng giá

Phạm vi bài toán

- Chỉ xét một cổ phiếu duy nhất: Google (GOOGL)
- Dữ liệu theo tần suất ngày, trong khoảng 10/10/2024 – 10/10/2025
- Tin tức được thu thập từ GDELT, sentiment gồm Positive / Neutral / Negative
- Không xét yếu tố vĩ mô (lãi suất, chỉ số thị trường, chính sách kinh tế...)

ĐỐI TƯỢNG VÀ CÔNG CỤ SỬ DỤNG



Đối tượng nghiên cứu

- Dữ liệu giá cổ phiếu Google (GOOGL) theo ngày
- Dữ liệu tin tức liên quan đến Google thu thập từ GDELT
- Xu hướng giá cổ phiếu được mô hình hóa dưới dạng bài toán phân loại nhị phân (Tăng / Giảm)

Công cụ & Công nghệ

- Ngôn ngữ: Python
- Xử lý dữ liệu: Pandas, NumPy
- Machine Learning: Xg Boost, Scikit-learn (Random Forest)
- NLP & Sentiment Analysis: FinBERT
- Trực quan & đánh giá: Matplotlib, Confusion Matrix, ROC Curve
- Môi trường phát triển: VS Code, Conda

NGUỒN DỮ LIỆU SỬ DỤNG

- Giá cổ phiếu: Yahoo Finance
 - Cổ phiếu Google (GOOGL)
 - Open, High, Low, Close, Volume
 - Theo ngày (10/10/2024 – 10/10/2025)
- Tin tức: GDELT Project
 - Tin liên quan đến Google
 - 2,448 bản tin
 - Phân tích cảm xúc bằng FinBERT
 - Nhãn: Positive / Neutral / Negative
- Dataset huấn luyện:
 - 214 mẫu sau khi merge & làm sạch dữ liệu

THÁCH THỨC KỸ THUẬT

- Dữ liệu không đồng bộ: Giá theo ngày, tin tức không đều
- Tin tức nhiễu: Không phải tin nào cũng ảnh hưởng giá
- Dữ liệu nhỏ: ~214 mẫu → nguy cơ overfitting
- Data leakage: Cần loại bỏ đặc trưng chứa thông tin tương lai
- Time-series split: Không được shuffle khi train/test

KIẾN TRÚC TỔNG THỂ PIPELINE

Step 1 – Tiền xử lý dữ liệu

Python xử lý dữ liệu giá cổ phiếu và tin tức, làm sạch, đồng bộ theo ngày và tạo tập dữ liệu đặc trưng.

Step 2 – Phân tích cảm xúc

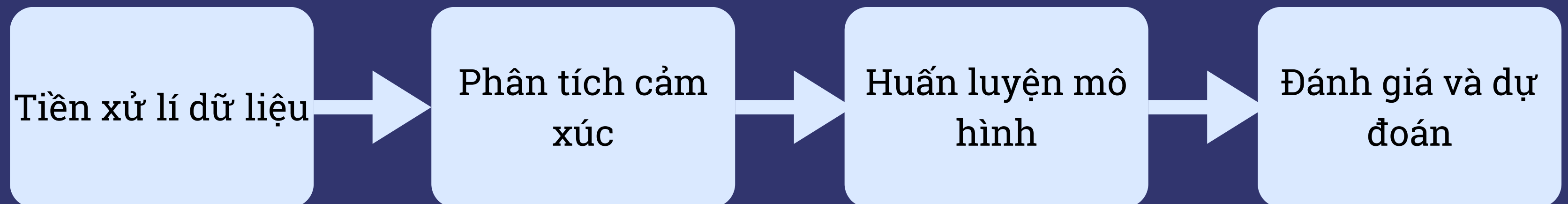
Áp dụng mô hình FinBERT để gán nhãn sentiment (Positive / Neutral / Negative) cho tin tức.

Step 3 – Huấn luyện mô hình

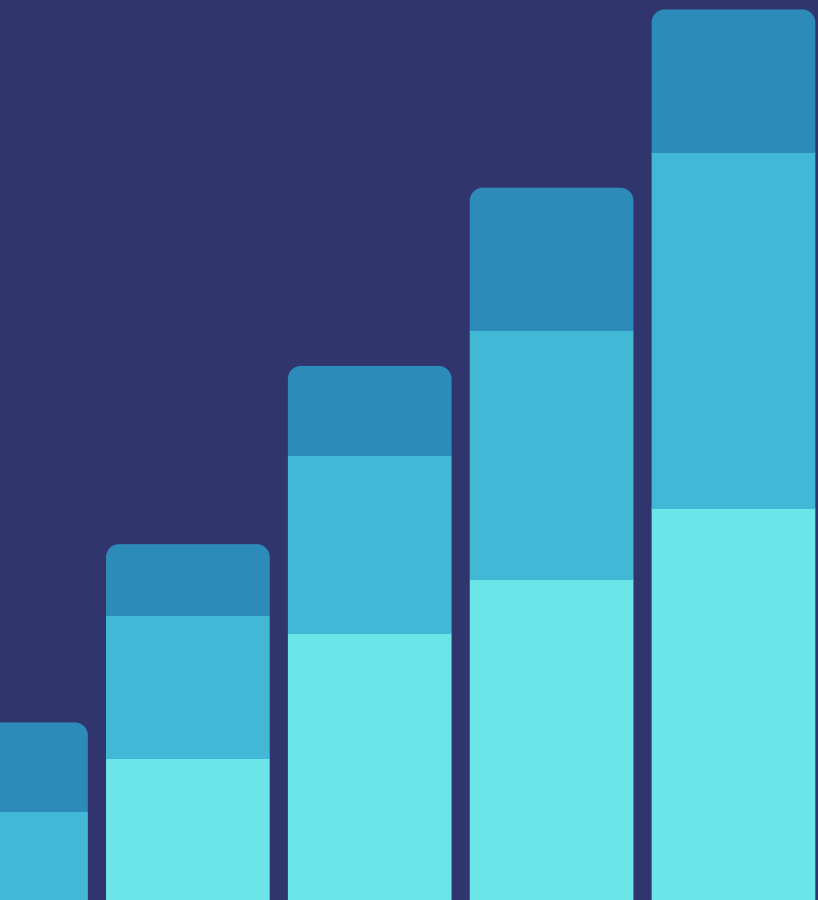
Kết hợp dữ liệu giá và sentiment, huấn luyện mô hình Xg Boost để dự đoán xu hướng giá.

Step 4 – Đánh giá & dự đoán

Đánh giá mô hình bằng các chỉ số chuẩn và thực hiện dự đoán xu hướng cho ngày giao dịch tiếp theo.



QUY TRÌNH TIỀN XỬ LÝ DỮ LIỆU



- Thu thập giá cổ phiếu và tin tức
- Chuẩn hóa mốc thời gian theo ngày
- Làm sạch dữ liệu, xử lý missing values
- Phân tích cảm xúc tin tức bằng FinBERT
- Gộp dữ liệu news – price theo ngày
- Loại bỏ các đặc trưng gây data leakage

Kết quả: một tập dữ liệu sạch, thống nhất, phù hợp để đưa vào pipeline.

DỮ LIỆU SỬ DỤNG



Giá cổ phiếu (GOOGL)

- Open, High, Low, Close, Volume
- Theo ngày – từ 10/10/2024 đến 10/10/2025

Tin tức tài chính

- Tiêu đề và nội dung tin liên quan đến Google
- Phân tích cảm xúc bằng FinBERT

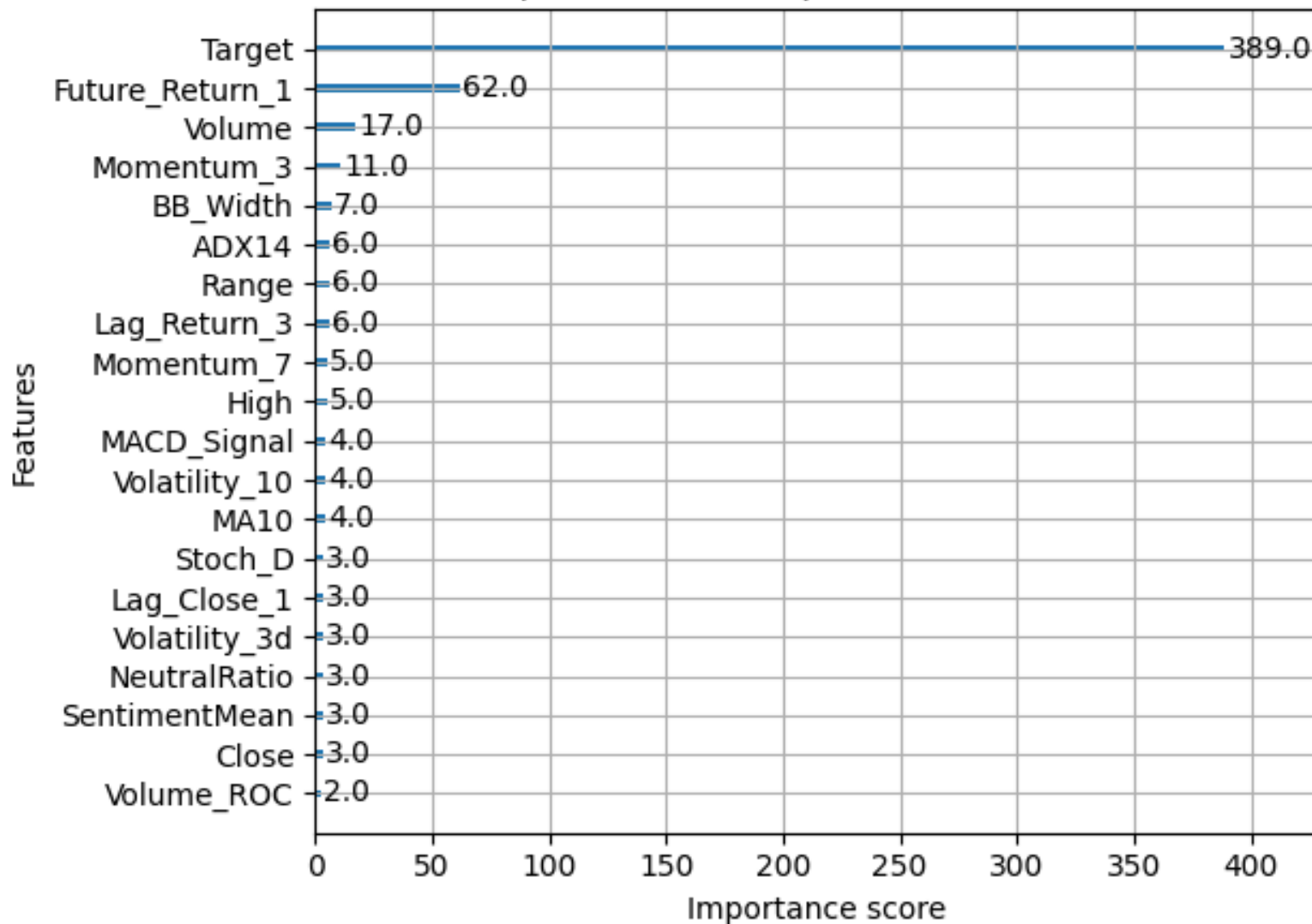
Quy mô dữ liệu

- Tin tức: 2,448 bản tin
- Giá cổ phiếu: 252 ngày giao dịch
- Dataset huấn luyện: 214 mẫu (sau xử lý)



FEATURE ENGINEERING

Top-20 Feature Importance (XGBoost)



- Đặc trưng giá:
- Return, Moving Average, Volatility
- Lag features:
- Giá, lợi suất và sentiment các ngày trước
- Candlestick features:
- Body, Range, Body/Range ratio
- Đặc trưng cảm xúc:
- Sentiment score tổng hợp từ tin tức

“Sentiment từ tin tức được tích hợp như một đặc trưng bổ sung cho mô hình.”



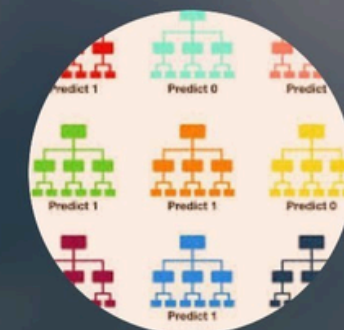
MÔ HÌNH SỬ DỤNG

- Random Forest
 - Mô hình tham chiếu để so sánh hiệu quả
 - Ổn định, dễ triển khai với dữ liệu bảng
- XGBoost (Mô hình chính)
 - Xử lý tốt dữ liệu tabular
 - Giảm overfitting nhờ regularization
 - Hiệu năng cao với feature phức tạp

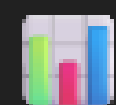
→ 📌 XGBoost được lựa chọn làm mô hình chính cho pipeline dự đoán.

XGBoost

RANDOM FOREST ALGORITHM

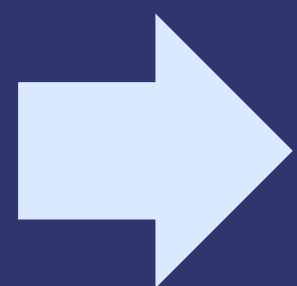


KẾT QUẢ MÔ HÌNH



Kết Quả Mô Hình (Test Set)

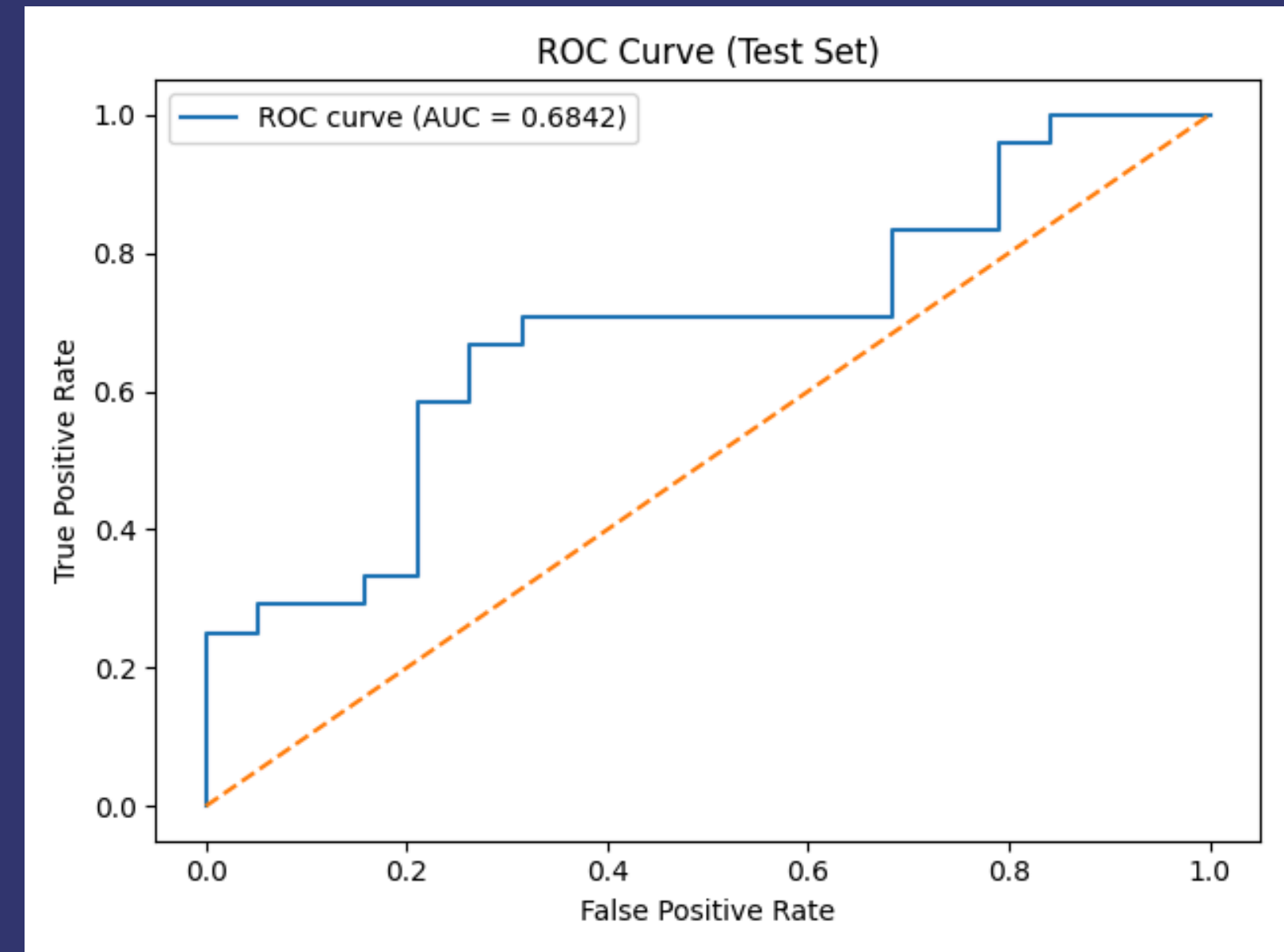
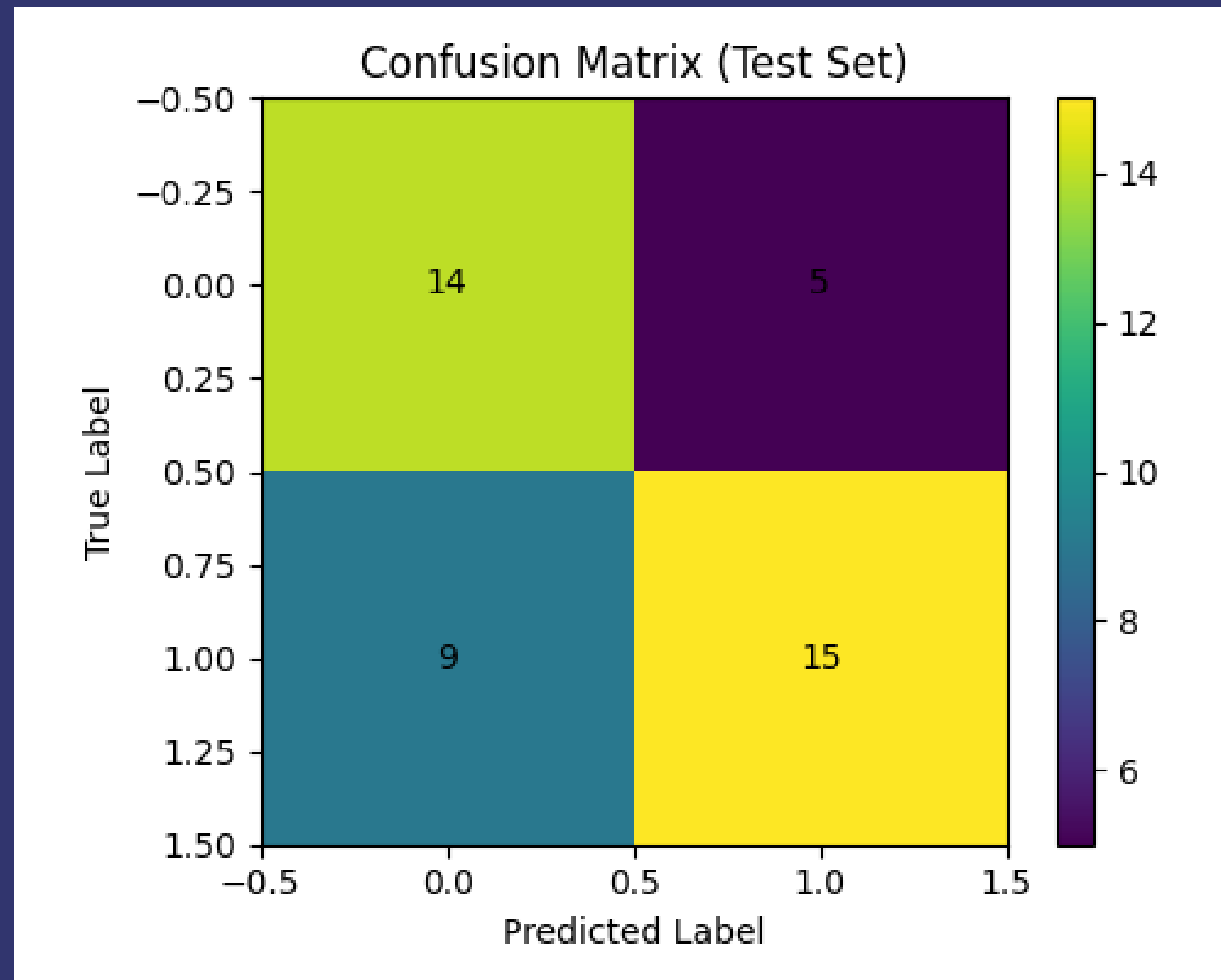
Mô hình	Accuracy	Precision	Recall	F1-score
Random Forest	0.5349	0.60	0.53	0.50
XGBoost ★	0.6744	0.75	0.63	0.68

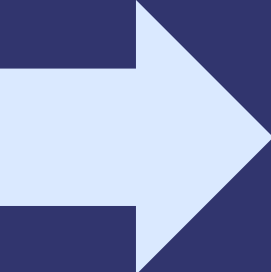


- Các chỉ số Precision / Recall / F1-score lấy theo weighted average
- XGBoost được highlight do cho kết quả vượt trội trên mọi chỉ số



CONFUSION MATRIX/ROC CURVE



- 
- Confusion Matrix cho thấy số lượng dự đoán đúng và sai của mô hình trên từng lớp Tăng/Giảm.
 - ROC Curve phản ánh khả năng phân biệt giữa hai lớp ở các ngưỡng khác nhau.



DEMO DỰ ĐOÁN THỰC TẾ

Input:

- Dữ liệu 1 ngày giao dịch gần nhất của Google
- Bao gồm:
 - Giá cổ phiếu (Open, High, Low, Close, Volume)
 - Chỉ báo kỹ thuật (MA, RSI, Volatility, Momentum,...)
 - Đặc trưng tin tức (Sentiment, News Count, Lag features)

Mô hình sử dụng:

- XGBoost (Time-series + Sentiment features)

Output dự đoán:

- 📈 Xu hướng ngày kế tiếp: TĂNG
- Xác suất tăng: 54.75%
- Xác suất giảm: 45.25%



DEMO CODE

NHẬN XÉT KẾT QUẢ

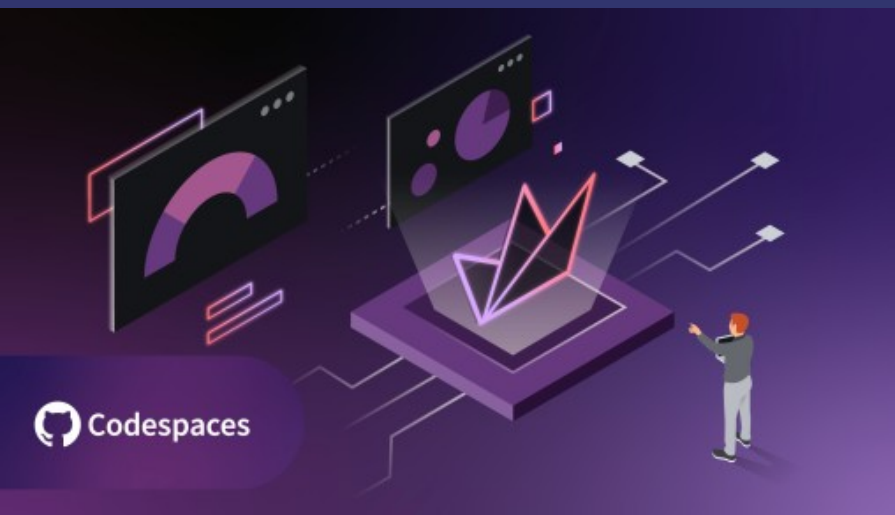
- XGBoost cho kết quả tốt
 - Accuracy test đạt ~67%
 - F1-score ~0.68, cân bằng giữa Precision và Recall
- Sentiment từ tin tức giúp cải thiện hiệu năng
- Mô hình ổn định trên tập test
- Có khả năng ứng dụng thực tế
 - Dự đoán được xu hướng ngày kế tiếp
 - Có thể mở rộng cho nhiều cổ phiếu khác

🔥 Best CV accuracy: 0.6000

📊 Classification Report (Best Fold Test Set):

	precision	recall	f1-score	support
0.0	0.5500	0.6875	0.6111	16
1.0	0.6667	0.5263	0.5882	19
accuracy			0.6000	35
macro avg	0.6083	0.6069	0.5997	35
weighted avg	0.6133	0.6000	0.5987	35

HẠN CHẾ



- Dữ liệu tin tức còn hạn chế, chưa bao phủ đầy đủ các nguồn và tin đột xuất
- Chưa xử lý các sự kiện bất thường (black swan) nên hiệu quả có thể giảm khi thị trường biến động mạnh
- Chưa tối ưu sâu hyperparameter, mô hình vẫn còn dư địa cải thiện

“Các hạn chế này cho thấy mô hình chưa hoàn hảo, nhưng hoàn toàn có khả năng cải thiện nếu mở rộng dữ liệu và tối ưu sâu hơn.”

KẾT LUẬN VÀ HƯỚNG PHÁT

- Kết luận
 - Xây dựng thành công pipeline Machine Learning hoàn chỉnh cho bài toán dự đoán xu hướng giá
 - Việc kết hợp dữ liệu giá và tin tức cho thấy là một hướng tiếp cận khả thi
- Hướng phát triển
 - Ứng dụng Deep Learning (LSTM, Transformer) cho dữ liệu chuỗi thời gian
 - Mở rộng và đa dạng hóa nguồn tin tức tài chính
 - Mở rộng bài toán sang dự đoán đa lớp hoặc hồi quy (regression)



THANK YOU