DATA EXPROLERS 2025

PROJECT: DỰ ĐOÁN GIÁ CHỨNG KHOÁN

Team HoiQuy

Vòng 2

ĐẶT VẤN ĐỀ

BỐI CẢNH

Thị trường tài chính ngày càng trở nên gần gũi và thu hút lượng lớn các nhà đầu tư mới, từ đó dẫn đến sự tăng trưởng chóng mặt về dữ liệu tài chính và nhu cầu dự báo chính xác từ bộ dữ liệu để đưa ra các quyết định

Vấn đề đặt ra

Dự báo giá cổ phiếu là một thách thức do biến động phức tạp của thị trường. Khối lượng dữ liệu khổng lồ: Bao gồm dữ liệu số (giá, giao dịch) và dữ liệu phi cấu trúc (báo cáo, tin tức). Con người khó xử lý kịp thời và toàn diện

Cơ hội

Sự phát triển của trí tuệ nhân tạo (AI) và mô hình ngôn ngữ lớn (LLM) mở ra khả năng khai thác dữ liệu hiệu quả hơn, là giải pháp hỗ trợ việc ra quyết định

MỤC TIÊU CỦA BÁO CÁO:

Ứng dụng các công nghệ trí tuệ nhân tạo hiện đại nhằm nâng cao khả năng phân tích và dự báo thị trường tài chính, đặc biệt trong việc khai thác dữ liệu và dự báo giá cổ phiếu.

TỔNG QUAN VỀ HAI MÃ CỔ PHIẾU FPT VÀ CMG

FPT (mã FPT): Tập đoàn công nghệ hàng đầu Việt Nam, hoạt động toàn diện từ phần mềm đến viễn thông.

CMC (mã CMG): Công ty công nghệ quy mô vừa, nổi bật trong an ninh mạng và hạ tầng.

FPT & CMG đều thuộc ngành công nghệ, bị ảnh hưởng bởi yếu tố tài chính – ngành – vĩ mô. Thêm vào đó, 2 công ty đều có nguồn dữ liệu đa dạng:

- Dữ liệu có cấu trúc: giá cổ phiếu, giao dịch, tài chính.
- Dữ liệu phi cấu trúc: báo cáo, tin tức, thảo luận.

Tạo cơ hội áp dụng mô hình AI khai thác sâu dữ liệu

TIỀN XỬ LÝ DỮ LIỆU

Quy trình tổng quát

1. Thu thập và chuẩn hóa dữ liệu.
2. Trích xuất đặc trưng từ dữ liệu số và văn bản.
3. Huấn luyện mô hình học máy.
4. Tích hợp mô hình RAG và LLM.
5. Dự báo giá và đánh giá kết quả.

CAFEF

VIETSTOCK refresh always

VnEconomy

- **Thời gian thu thập:** 1/2023 – 12/3/2025.
- **Nguồn dữ liệu số:** giá cổ phiếu, giao dịch, dòng tiền, v.v. (Investing, Vietstock, FireAnt, Bloomberg,...)
- **Nguồn dữ liệu văn bản**: báo cáo tài chính, tin tức, thảo luận (Cafef, VnEconomy, báo cáo doanh nghiệp,..)

Các nhóm dữ liệu chính:

- Dữ liệu ngành: dòng tiền, hiệu suất ngành, giao dịch nội bộ.
- Dữ liệu giá & khối lượng FPT, CMG.
- Dữ liệu tin tức.
- Báo cáo tài chính & chỉ số định lượng.

ĐÁNH GIÁ CHẤT LƯỢNG DỮ LIỆU VÀ XỬ LÝ DỮ LIỆU

VẤN ĐỀ GẶP PHẢI

- Dữ liệu không đầy đủ theo thời gian
- Tồn tại nhiều giá trị ngoại lai
- Thiếu feature phản ánh ngữ cảnh
- Không đồng nhất về định dạng và tên biến giữa các file
- Mất cân bằng trong phân phối dữ liệu
- Thiếu cột định danh (label) rõ ràng cho một số tập dữ liệu

CÁCH XỬ LÝ

Xử lý: chuẩn hóa, Denton, trích xuất đặc trưng, vector hóa

- Chuẩn hóa thời gian, dữ liệu giá: biến đổi log, tính trung bình trượt (MA).
- Kết hợp đa nguồn (structured + unstructured).
- Loại bỏ nhiễu, Chuyển đổi văn bản thành embedding (TF-IDF, BERT, etc.).
- Tổng hợp thành bộ đặc trưng đầu vào cho mô hình.

ĐỊNH DẠNG DỮ LIỆU

- **Xử lý giá trị ngoại lai:** Giữ lại ngoại lai nhưng sử dụng các scaler mạnh mẽ.
- **Xử lý mất cân bằng phân phối:** Sử dụng RobustScaler, PowerTransformer.

PHÂN CHIA TẬP DỮ LIỆU

Chia dữ liệu thành tập huấn luyện (train), tập kiểm định (validation), và tập kiểm tra (test) theo thứ tự thời gian

XÂY DỰNG MÔ HÌNH VÀ ĐÁNH GIÁ, CẢI TIẾN

BASELINE - LINEAR REGRESSION

Dùng làm mô hình tham chiếu ban đầu, giả định quan hệ tuyến tính đơn giản

TREE-BASED MODELS

- **Random Forest**: Khả năng bắt các mối quan hệ phi tuyến, kháng nhiễu tốt.
- **XGBoost:** Tối ưu hoá tốt, hiệu suất cao trong bài toán hồi quy.
- **LightGBM**: Sử dụng ít bộ nhớ, hỗ trợ dữ liệu lớn. Hỗ trợ tính năng quan trọng như xử lý missing values, tự động chọn split.

DEEP LEARNING – LSTM

- Long Short-Term Memory (LSTM): Phù hợp với dữ liệu chuỗi thời gian.
- LSTM xử lý mối quan hệ thời gian trong dữ liệu giá cổ phiếu quá khứ.

CÁC MÔ HÌNH THÀNH PHẦN

M1_FinInd: Dự đoán giá dựa trên BCTC & Dữ liệu Ngành

M2_MktFI: Dự đoán giá dựa trên Dữ liệu Thị trường & Đầu tư Nước ngoài

M3_Ensemble: Mô hình Tổng hợp (Ensemble) Tối ưu hóa Dự đoán Giá Ngắn hạn (T+1)

M4_LSTM_Forecast: Mô hình Mạng LSTM Dự báo Giá Cổ phiếu Đa bước (3 ngày)

Quy trình chung bao gồm chuẩn bị dữ liệu, phân chia tập, huấn luyện từng mô hình (M1, M2, M3, M4) và quy trình RAG_News, tối ưu siêu tham số dựa trên tập validation, và sử dụng các kỹ thuật như:

- Early Stopping
- ModelCheckpoint
- ReduceLROnPlateau.

ĐỂ NÂNG CAO CHẤT LƯỢNG MÔ HÌNH, CÓ THỂ TÍCH HỢP THÊM RAG VÀ LLM

01 RAG

Mô hình kết hợp truy xuất thông tin (retrieval) và tạo văn bản (generation) để tạo phản hồi chính xác và có ngữ cảnh.

RAG = Tìm kiếm văn bản liên quan trong corpus tài chính
→ Đưa vào LLM để sinh câu trả lời có dẫn chứng thực tế

Ý nghĩa

- Giúp mô hình dự báo hiểu bối cảnh thị trường tại thời điểm cụ thể.
- Là cầu nối giữa dữ liệu phi cấu trúc (text) và mô hình dự báo định lượng.

02 LLM

- ChatGPT (GPT-4): Ưu điểm tổng quát, hiểu tốt tiếng Việt.
- LLaMA 2.3: Nhẹ, mã nguồn mở, dễ tùy chỉnh.

Ý nghĩa

- Tự động hóa quy trình phân tích văn bản.
- Hỗ trợ nhà đầu tư hiểu nhanh thông tin định tính, gắn vào mô hình định lượng.

```
Gửi prompt RAG hoàn chỉnh đến LLM...
Đang yêu cầu llama3.2:latest tạo câu trả lời RAG...
Thời gian tạo sinh của Ollama: 14.91 giây

Câu trả lời từ hệ thống RAG:
FPT trả cổ tức lần cuối vào ngày 13/12/2024 (Friday, Tháng December/2,024, Quý 4).
```

```
Gửi prompt RAG hoàn chỉnh đến LLM...
Đang yêu cầu llama3.2:latest tạo câu trả lời RAG...
Thời gian tạo sinh của Ollama: 6.81 giây

Câu trả lời từ hệ thống RAG:
Giá cổ phiếu VCB hiện tại không được cung cấp trong dữ liệu được cung cấp.
```

Demo hoạt động RAG khi có dữ liệu và không có dữ liệu

XÂY DỰNG MÔ HÌNH VÀ ĐÁNH GIÁ, CẢI TIẾN

Mô hình / Cổ phiếu	MSE (Giá)	MAE (Giá)	RMSE (Giá)
M1_FinInd (CMG)	7,409,853	1,829.70	2,722.1
M1_FinInd (FPT)	299,436,460	15,747.80	17,304.2

Mô hình / Cổ phiếu	MSE (Giá)	MAE (Giá)	RMSE (Giá)
M2_MktFI (CMG)	935,179	6.833	967
M2_MktFI (FPT)	115,311,214	9,676.60	10,738.3

Mô hình / Cổ phiếu	MSE (Giá)	MAE (Giá)	RMSE (Giá)
M3_Ensemble (CMG)	2,243,672.1	1,263.7	1,497.8
M3_Ensemble (FPT)	66,048,799.7	6,011.2	8127

Mô hình / Cổ phiếu	MAE (3 ngày)	Sai lệch từng ngày
M4_LSTM_Forecast (CMG)	~114 VND	Ngày 1: 14.5 VNĐ; Ngày 2: 104.7 VNĐ; Ngày 3: 107.1 VNĐ
M4_LSTM_Forecast (FPT)	~3,861 VND	Ngày 1: 1,430 VNĐ; Ngày 2: 6,290 VNĐ; Ngày 3: 7,462 VNĐ

TỔNG HỢP KẾT QUẢ

Phân tích và diễn giải kết quả tổng thể

- Hiệu suất khác biệt (CMG vs FPT): CMG dễ dự đoán hơn FPT bằng các mô hình cơ sở.
- Vai trò M3_Ensemble và RAG_News: M3 nâng cao hiệu suất, đặc biệt cho FPT, nhờ features từ RAG.
- Phụ thuộc của M4 vào M3: Chất lượng dự báo T+1 của M3 ảnh hưởng trực tiếp đến M4.
- Tác động Features Tin tức: Cần phân tích sâu hơn đóng góp của các features từ RAG_News.

Hệ thống Đa mô hình cho kết quả khá tốt, nhất là trong việc bắt xu hướng ngắn hạn và dự đoán các mã cổ phiếu của công ty nhỏ ít biến động như CMC. Tuy nhiên, độ chính xác còn bị giới hạn do thiếu dữ liệu cảm tính (news, sự kiện) và không phản ánh ngay các yếu tố phi cấu trúc

Mô hình	Ưu điểm	Hạn chế
ML cơ bản	Đơn giản, nhanh	Thiếu ngữ cảnh
ML + LLM	Giải thích tốt, giàu ngữ nghĩa	Tốn tài nguyên
RAG + LLM	Linh hoạt, cập nhật mới	Cần hạ tầng mạnh

=> Mô hình tích hợp RAG + LLM có tiềm năng mở rộng mạnh.

TỔNG KẾT DỰ ÁN

HẠN CHẾ

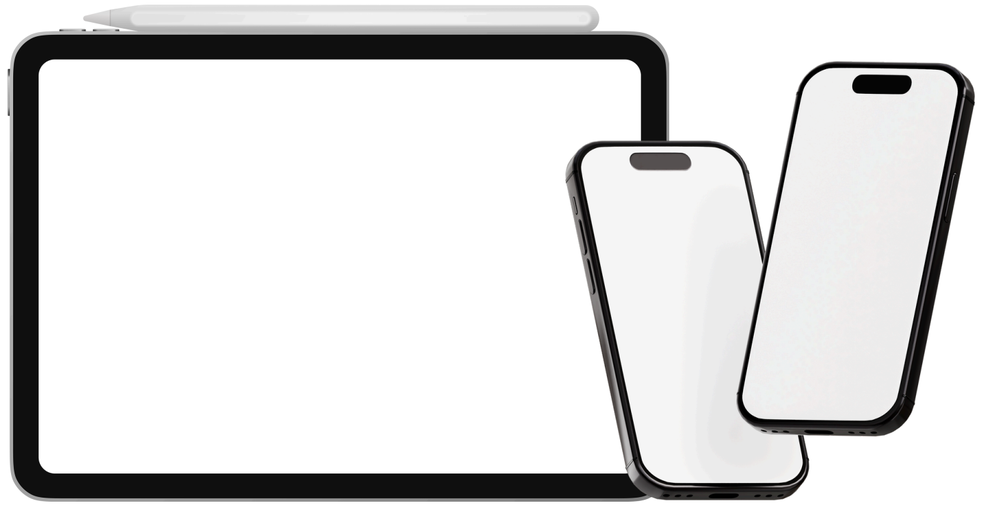
- Thiếu dữ liệu chất lượng cao, thiếu dữ liệu cảm xúc NĐT, khó xác định nhân quả từ tin tức
- LLM chưa hoàn toàn tối ưu cho tiếng Việt
- Cách trả lời chưa đa dạng

ĐỀ XUẤT CẢI TIẾN

- Bổ sung dữ liệu kinh tế vĩ mô: Lãi suất, CPI, GDP để nâng cao độ bao phủ.
- Áp dụng thêm các kỹ thuật phân tích cảm xúc (sentiment analysis)
- Mở rộng dự báo dài hạn: Từ vài ngày → vài tuần/tháng để hỗ trợ nhà đầu tư dài hạn.
- Xây dựng app hoặc web app: Giao diện đơn giản cho phép nhập mã cổ phiếu → trả về dự báo + phân tích ngữ cảnh tự động bằng AI.

TIỀM NĂNG PHÁT TRIỂN

- Áp dụng cho các cổ phiếu khác trong ngành công nghệ như HTP, ITD,...
- Tùy chỉnh logic và prompt để mở rộng sang ngành khác (ngân hàng, bất động sản...).
- Có thể xây dựng thành công cụ hỗ trợ đầu tư thông minh hoặc hệ thống phân tích tài chính tự động.



Thank you for your attention

Team HoiQuy

Vòng 2