Phân tích, dự đoán giá chứng khoản bằng mô hình thống kê, máy học và học sâu

Trần Thị Mỹ Xoan - 21522815
Lê Anh Tuấn Dũng - 21521974
Lê Thị Ánh Hồng - 21520245
Nguyễn Thị Mai Liên - 21522283
Đỗ Sĩ Đạt - 21521932

Ngày 9 tháng 6 năm 2024

Tóm tắt nội dung: Khi thực hiện đầu tư vào các cổ phiếu chứng khoán, việc sử dụng các mô hình dự đoán giá cổ phiếu trước khi thực hiện một giao dịch đầu tư có thể giúp cho ta có các quyết định đúng đắn hơn khi ra quyết định thực hiện 1 giao dịch. Bài báo này trình bày phương pháp dự đoán giá cổ phiếu của ba công ty lớn Samsung, LG và Sony. Chúng tôi sẽ thực hiện dự đoán bằng cách sử dụng các mô hình thống kê, học máy, học sâu: VARMA, XGBoost, NBeats, Gradient Boosting, LightGBM. Bộ dữ liệu được sử dụng lấy từ ngày 1/3/2019 đến ngày 1/6/2024. Sau khi thực hiện dự đoán thì ta sẽ thực hiện sử dụng các độ đo MAPE, MSE, RMSE để đánh giá các mô hình. Cuối cùng, sử dụng các mô hình để thực hiện dự đoán giá chứng khoán cho 30, 60 và 90 ngày tiếp theo.

Từ khoá: VARMA, XGBoost, NBeats, Gradient Boosting, LightGBM. Mô hình thống kê, mô hình học máy, mô hình học sâu, dự đoán giá chứng khoán.

1 GIỚI THIỆU

2 NGHIÊN CỨU LIÊN QUAN

[4] Mô hình VARMA thường được sử dụng để dự báo dữ liệu chuỗi thời gian đa biến. Mô hình VARMA là sự mở rộng của mô hình ARMA trong chuỗi thời gian đơn biến (Lütkepohl, 2005; Wei, 1990) và được sử dụng với điều kiện rằng dữ liệu phải ổn định theo thời gian (Lütkepohl, 2005). Mô hình VARMA (p,q) là sự kết hợp của mô hình VAR (p) và mô hình vector trung bình động (q) (VMA (q)). Bài báo này xác định 4 mô hình VARMA (p,q) lần lượt là (1,1), (2, 1), (3, 1), (4, 1), Việc lựa chọn mô hình tốt nhất được thực hiện bằng cách sử dụng một số tiêu chí thông tin (AICC, HQC, AIC, và SBC). Các giá trị nhỏ nhất của những tiêu chí này cho biết mô hình tốt nhất. Mô hình phù hợp nhất được chọn là VARMA(2, 1). Kết quả dự báo cho thấy sai số chuẩn tăng lên theo thời gian; sai số chuẩn trong tháng đầu tiên tương đối nhỏ so với dự đoán trung bình, nhưng tăng lên theo thời gian đến dự báo cho 12 tháng tiếp theo. Điều này cho thấy mô hình là hợp lý khi dự báo cho các khoảng thời gian ngắn, nhưng kết quả không ổn định (do sai số chuẩn cao hơn) khi dự báo cho các khoảng thời gian dài.

[5] Và để tìm ra tham số tối ưu cho mô hình VARMA, ta sử dụng `auto_arima` để thực hiện tìm các tham số tối ưu (p, q) cho mỗi biến. Cặp tham số

của một biến có AIC nhỏ nhất là phù hợp nhất. Sau đó, huấn luyện với các cặp tham số (p, q) tìm được. Sau khi chạy, ta sử dụng chỉ số đánh giá RMSE để xem cặp tham số (p, q) nào phù hợp nhất với mô hình. RMSE càng nhỏ thì cặp tham số đó càng phù hợp.

3 PHƯƠNG PHÁP

3.1 DATASET

3.2 CHỈ SỐ ĐÁNH GIÁ MÔ HÌNH

3.3 THUẬT TOÁN RNN

3.4 THUẬT TOÁN LIGHTGBM

3.5 THUẬT TOÁN VARMA

Mô hình VARMA thường được sử dụng để dự báo dữ liệu chuỗi thời gian đa biến . Mô hình VARMA là một mở rộng của mô hình ARMA trong chuỗi thời gian đơn biến (Lutkepohl, 2005; Wei, 1990) và được sử dụng với điều kiện dữ liệu phải là dừng theo thời gian (Lutkepohl, 2005). Mô hình VARMA (p,q) là sự kết hợp của mô hình VAR (p) và mô hình trung bình trượt vector (q) (VMA (q)).

- Công thức:

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{y}_{t-i} + \sum_{j=1}^{q} \mathbf{\Theta}_j \boldsymbol{\epsilon}_{t-j} + \boldsymbol{\epsilon}_t \qquad (1)$$

Trong đó:

- $\mathbf{y}_t$ là vector của các biến tại thời điểm t.
- $\mathbf{c}$ là một vector hằng số.
- $\mathbf{\Phi}_i$ là ma trận hệ số của chuỗi thời gian (độ trễ, lag) i của mô hình AR.
- $\mathbf{\Theta}_j$ là ma trận hệ số của chuỗi thời gian (độ trễ, lag) j của mô hình MA.
- $\boldsymbol{\epsilon}_t$ là vector của các nhiễu ngẫu nhiên tại thời điểm t.

3.6 THUẬT TOÁN GRADIENT BOOSTING REGRESSION

3.7 THUẬT TOÁN XGBoost

3.8 THUẬT TOÁN NBEAT

4 THỰC NGHIỆM

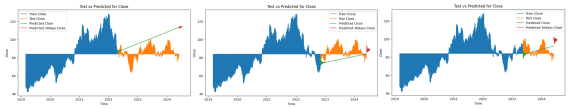
4.1 MÔ HÌNH DỰ ĐOÁN

4.1.1 VARMA

Hình 1: LG 30 DAYS 6:4, 7:3, 8:2

Hình 2: LG 60 DAYS 6:4, 7:3, 8:2

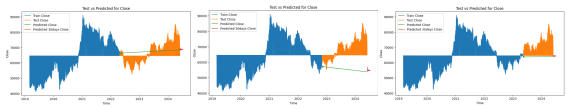
Hình 3: LG 90 DAYS 6:4, 7:3, 8:2



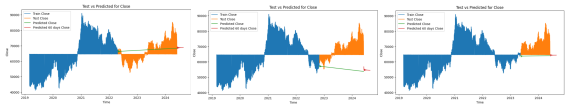
Hình 4: SONY 30 DAYS 6:4, 7:3, 8:2

Hình 5: SONY 60 DAYS 6:4, 7:3, 8:2

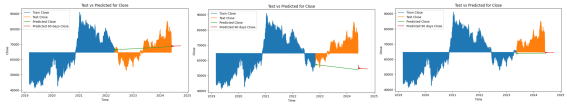
Hình 6: SONY 90 DAYS 6:4, 7:3, 8:2



Hình 7: SAMSUNG 30 DAYS 6:4, 7:3, 8:2



Hình 8: SAMSUNG 60 DAYS 6:4, 7:3, 8:2



Hình 9: SAMSUNG 90 DAYS 6:4, 7:3, 8:2

4.1.2 GRADIENT BOOSTING

4.1.3 XGBOOST

4.2 THANG ĐO VÀ KẾT QUẢ

4.2.1 ĐÁNH GIÁ MÔ HÌNH VỚI DATASET LG

4.2.2 ĐÁNH GIÁ MÔ HÌNH VỚI DATASET SONY

4.2.3 ĐÁNH GIÁ MÔ HÌNH VỚI DATASET SAMSUNG

5 TÀI LIỆU THAM KHẢO

[4]: Warsono, Edwin Russel, Wamiliana*, Widiarti, Mustofa Usman, "Modeling and Forecasting by the Vector Autoregressive Moving Average Model for Export of Coal and Oil Data (Case Study from Indonesia over the Years 2002-2017)". International Journal of Energy Economics and Policy, 2019

[5]: Yugesh Verma, "A Guide to VARMA with Auto ARIMA in Time Series Modelling". analyticsindiamag.com, 2021