ỨNG DỤNG THUẬT TOÁN RỪNG NGẪU NHIÊN TRONG   
DỰ ĐOÁN GIÁ CỔ PHIẾU APPLE

Phạm Hoàng Phúc  
Châu Mỹ Uyên

**Tóm tắt**

Apple là một trong những công ty công nghệ hàng đầu và mạng xã hội phổ biến nhất trên toàn cầu, với ảnh hưởng lớn đến thị trường tài chính và cổ phiếu. Vì vậy, việc dự đoán giá cổ phiếu Apple có thể mang lại nhiều lợi ích cho nhà đầu tư và những người quan tâm đến thị trường tài chính. Báo cáo này vận dụng mô hình trí tuệ nhân tạo Random Forest để dự báo giá cổ phiếu thế giới dựa trên ngày giao dịch, giá mở cửa, giá đóng cửa, giá đóng cửa có điều chỉnh, giá cổ phiếu cao nhất, giá thấp nhất và khối lượng giao dịch của cổ phiếu Apple trong ngày. Kết quả nghiên cứu cho thấy mô hình Random Forest có thể dự báo tốt giá cổ phiếu Apple. Kết quả dự báo từ tập huấn luyện và tập kiểm tra bám sát giá thực tế và có sai số dự báo ở mức thấp, với sai số trung bình tuyệt đối lần lượt là MAE = 0.29 và 1.38; và căn bậc hai trung bình bình phương sai số dự báo tương ứng là RMSE = 0.421 và 1.76.

**Từ khóa**: Random forest, Stock Prediction, cổ phiếu Apple.

1. **Giới thiệu**

Các công nghệ tiên tiến như trí tuệ nhân tạo, học máy, khoa học dữ liệu, dữ liệu lớn,… đã được nghiên cứu và áp dụng để giải quyết các vấn đề trong nhiều lĩnh vực. Báo cáo này tập trung mô tả về thuật toán phân loại Random Forest trong học máy và giải thích trường hợp sử dụng nó trong lĩnh vực tài chính. Cụ thể là, chúng tôi sử dụng Random Forests để dự đoán giá cổ phiếu vì khả năng xử lý dữ liệu lớn và phức tạp, khả năng xử lý các tình huống đa tầng và phi tuyến, cũng như tính ổn định và kháng nhiễu của mô hình. Với sự hỗ trợ của Random Forests, chúng tôi tin rằng chúng tôi có thể cung cấp các dự đoán chính xác và tin cậy về giá cổ phiếu, giúp nhà đầu tư và những người quan tâm có thông tin quan trọng để ra quyết định đúng đắn.

Việc chọn tập dữ liệu Apple để dự đoán giá cổ phiếu bằng thuật toán Random forests là vì Apple là một trong những công ty công nghệ hàng đầu và mạng xã hội phổ biến nhất trên toàn cầu, với ảnh hưởng lớn đến thị trường tài chính và cổ phiếu. Vì vậy, việc dự đoán giá cổ phiếu Apple có thể mang lại nhiều lợi ích cho nhà đầu tư và những người quan tâm đến thị trường tài chính.

Lựa chọn sử dụng thuật toán cây ngẫu nhiên (Random Forest) để dự đoán giá cổ phiếu Apple cũng có những lý do đáng chú ý. Cây ngẫu nhiên là một thuật toán học máy mạnh mẽ và phổ biến, đặc biệt thích hợp cho bài toán dự đoán trong lĩnh vực tài chính.

Sử dụng cây ngẫu nhiên trong dự đoán giá cổ phiếu Apple cũng mang lại sự linh hoạt và khả năng giải thích kết quả. Cây ngẫu nhiên có thể cung cấp thông tin về độ quan trọng của các biến đầu vào trong việc dự đoán, cho phép nhà đầu tư và các chuyên gia tài chính hiểu rõ hơn về yếu tố nào đóng vai trò quan trọng trong sự biến động của giá cổ phiếu Apple.

Với tập dữ liệu Apple và sử dụng thuật toán cây ngẫu nhiên, ta hy vọng có thể xây dựng một mô hình dự đoán giá cổ phiếu chính xác và tin cậy, đồng thời cung cấp cái nhìn sâu hơn về các yếu tố ảnh hưởng đến giá cổ phiếu Apple. Tuy nhiên, cần lưu ý rằng dự đoán giá cổ phiếu luôn mang tính dự đoán và có yếu tố không chắc chắn, do đó, việc tham khảo nhiều nguồn thông tin và phân tích cẩn thận vẫn là rất quan trọng trong quá trình đầu tư và ra quyết định.

1. **Thuật toán random forest**

RF được giới thiệu lần đầu tiên bởi Breiman vào năm 2001. Nó là một phương pháp phân loại và hồi quy theo phương thức học có kiểm soát, bao gồm sự kết hợp nhiều cây dự báo/cây ra quyết định. Mỗi cây ra quyết định được tạo ra dựa trên một tập mẫu huấn luyện ngẫu nhiên được tạo ra từ tập mẫu huấn luyện ban đầu với cùng độ lớn theo nguyên tắc lấy mẫu có hoàn lại (phương pháp Bootstrap) nghĩa là một mẫu có thể được lấy nhiều lần. Các cây ra quyết định dựa trên tập mẫu vừa tạo ra với nguyên tắc chỉ sử dụng một số lượng biến đầu vào tại mỗi nút phân chia. Kết quả cuối cùng là giá trị trung bình kết quả thu được từ tất cả các cây ra quyết định. Bằng cách sử dụng thật nhiều cây ra quyết định, sai số ước lượng của mô hình sẽ được giảm đi đáng kể. Nguyên lý của thuật toán được tóm tắt trong Hình 1.

A picture containing text, diagram, font, plan

Description automatically generated

Hình 1. Thuật toán rừng ngẫu nhiên

Nó hoạt động theo bốn bước:

1. Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.
2. Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.
3. Hãy bỏ phiếu cho mỗi kết quả dự đoán.
4. Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.

Sử dụng tập train con và cây quyết định độc lập giúp Random Forests có tính đa dạng và khả năng tổng quát hóa. Việc kết hợp các dự đoán từ các cây con giúp cân nhắc và làm giảm ảnh hưởng của nhiễu và overfitting, đồng thời cung cấp một dự đoán cuối cùng ổn định và tin cậy.

Trong bài toán hồi quy, các cây dự báo sẽ nhận các giá trị số cụ thể thay vì kỹ thuật phân loại (ra quyết định). Trong thiết kế các cây ra quyết định của phân tích hồi quy, các cây được cho phép phát triển đến độ sâu tối đa của dữ liệu huấn luyện mà không tiến hành giản lược lại (ngắt cành). Đây cũng là một ưu điểm chính của thuật toán này bởi vì việc thực hiện giản lược cây là một yếu tố chính ảnh hưởng đến hiệu suất của mô hình. Breiman cũng cho rằng khi số lượng cây tăng lên, sai số tổng quát luôn hội tụ ngay cả khi không giản lược cây, và việc xử lý hiện tượng mô hình quá khớp được thực hiện dựa trên nguyên tắc luật số lớn (Strong Law of Large Numbers). Số lượng biến được sử dụng (N) tại mỗi nút để tạo ra một cây ra quyết định và số lượng cây ra quyết định (k) được sử dụng là hai tham số do người dùng lựa chọn trước.

Dựa trên trình bày trên đây, chúng ta rút ra được các thông số quan trọng có thể ảnh hưởng đến hiệu suất của mô hình rừng ngẫu nhiên là: (1) số lượng cây ra quyết định được sử dụng; (2) kỹ thuật lấy mẫu (sử dụng hay không sử dụng kỹ thuật bootstrap); (3) số lượng biến được sử dụng tại mỗi nút; (4) đặc điểm của tập dữ liệu bao gồm thông số đầu vào và thông số đầu ra của mô hình.

1. **Xây dựng thuật toán random forests để dự đoán giá cổ phiếu của Apple**
   1. **Dữ liệu và các biến sử dụng trong mô hình**

Sử dụng tập dữ liệu chứa thông tin về giá cổ phiếu Apple từ năm 2019 đến năm 2022. Có 1008 dòng và 7 cột gồm: 1 biến phụ thuộc và biến độc lập

Biến phụ thuộc:

- Biến Adj Close: Giá đóng cửa điều chỉnh của cổ phiếu Apple trong một ngày cụ thể. Đây là giá cổ phiếu được điều chỉnh bởi các yếu tố như chia cổ tức, chia cổ phiếu, hay cấp quyền mua cổ phiếu.

Biến độc lập:

- Giá mở cửa (Open): Giá cổ phiếu Apple khi phiên giao dịch bắt đầu trong một ngày cụ thể.

- Giá thấp nhất (Low): Giá cổ phiếu Apple thấp nhất đạt được trong phiên giao dịch trong một ngày cụ thể.

- Giá cao nhất (High): Giá cổ phiếu Apple cao nhất đạt được trong phiên giao dịch trong một ngày cụ thể.

- Giá đóng cửa (Close): Giá cổ phiếu Apple khi phiên giao dịch kết thúc trong một ngày cụ thể.

- Khối lượng giao dịch (Volume): Số lượng cổ phiếu Apple được giao dịch trong phiên giao dịch trong một ngày cụ thể. Các biến này được sử dụng để xây dựng mô hình dự đoán giá cổ phiếu Apple (biến Adj Close) dựa trên thông tin về giá cổ phiếu trong quá khứ và khối lượng giao dịch tương ứng.

**A screen shot of numbers

Description automatically generated with low confidence**

**3.2 Trực quan và phân tích dữ liệu**

**-** Biểu đồ giá mở cửa (Open): Biểu đồ này cho thấy sự biến động của giá cổ phiếu Apple từng ngày. Ta có thể quan sát xu hướng tăng/giảm của giá trong một khoảng thời gian nhất định.

**A picture containing text, screenshot, line, font

Description automatically generated**Biểu đồ thể hiện độ biến động giá mở cổ phiếu Apple theo từng ngày

**-** Biểu đồ giá đóng cửa (Close): Biểu đồ này cho thấy sự biến động của giá cổ phiếu Apple khi đóng phiên giao dịch theo từng ngày. Ta có thể quan sát xu hướng tăng/giảm của giá trong một khoảng thời gian nhất định.

A picture containing text, screenshot, line, font

Description automatically generated

Biểu đồ thể hiện độ biến động giá mở cổ phiếu Apple theo từng ngàys

**-** Biểu đồ giá điều chỉnh đóng cửa (Adj Close): Biểu đồ này hiển thị giá cổ phiếu đã được điều chỉnh theo các yếu tố như chia cổ tức, phát hành cổ phiếu thưởng, chia cổ phiếu, v.v. Nó cho ta cái nhìn tổng quan về sự biến động của giá cổ phiếu Apple theo thời gian.

**A picture containing text, screenshot, font, line

Description automatically generated**Biểu đồ thể hiện độ biến động giá điều chỉnh đóng cổ phiếu Apple theo từng ngày

* Biểu đồ giá cổ phiếu cao nhất (High): biểu độ thể hiện giá cổ phiếu Apple cao nhất biến động theo từng ngày. Cho ta thấy được sự tăng/giảm trong một khoảng thời gian nhất định.

A picture containing text, screenshot, line, font

Description automatically generated

Biểu đồ thể hiện giá cổ phiếu Apple cao nhất theo ngày

* Biểu đồ giá thấp nhất (Low): biểu đồ thể hiện giá cổ phiếu thấp nhất trong phiên giao dịch theo từng ngày, cho thấy được sự biến động tăng/giảm trong một khoảng thời gian.

A picture containing text, screenshot, line, font

Description automatically generated  
Biểu đồ thể hiện giá thấp nhất trong phiên giao dịch theo ngày

**-** Biểu đồ khối lượng giao dịch (Volume): Biểu đồ này cho thấy sự biến động của khối lượng cổ phiếu được giao dịch trong một phiên giao dịch. Nó có thể cung cấp thông tin về sự quan tâm và hoạt động giao dịch của thị trường đối với cổ phiếu Apple.

A picture containing text, screenshot, plot, diagram

Description automatically generated  
Biểu đồ thể hiện độ biến động khối lượng giao dịch cổ phiếu Apple theo từng ngày

A picture containing text, screenshot, plot, diagram

Description automatically generated  
Biểu đồ thể hiện các biến trong tập dữ liệu

Từ các biểu đồ trên, ta có thể rút ra các nhận định về xu hướng, biến động, và sự tương quan giữa các yếu tố liên quan đến cổ phiếu Apple, như giá cổ phiếu, khối lượng giao dịch, tổng giá trị giao dịch và lợi nhuận hàng ngày.

**3.3 Áp dụng mô hình dự đoán**

* Phân chia tập dữ liệu. Tập dữ liệu ban đầu sẽ được phân chia thành 2 tập riêng biệt: (1) tập huấn luyện được sử dụng cho quá trình huấn luyện mô hình và (2) tập kiểm tra được sử dụng để đánh giá hiệu quả của mô hình. Ở đây, chúng tôi chia tập huấn luyện 80% có 807 dòng và 20% còn lại là tập kiếm tra là 201 dòng.

A picture containing text, screenshot, software, display

Description automatically generated

* Từ dữ liệu đã được phân chia ta dùng để đưa vào mô hình dự đoán giá cổ phiếu Apple

A screen shot of a computer code

Description automatically generated with low confidence

A picture containing text, plot, line, screenshot

Description automatically generated

Biểu đồ các Features và Label trong tập train và tập test

* Tính toán các chỉ số đánh giá để đo lường hiệu suất của mô hình dự đoán giá cổ phiếu.

+ MAE: Là sai số trung bình tuyệt đối (Mean Absolute Error - MAE) của mô hình trên tập huấn luyện và tập kiểm tra. Đây là chỉ số đánh giá đo lường mức độ chính xác trung bình của dự đoán so với giá trị thực tế, bằng cách lấy trung bình giá trị tuyệt đối của sai số.

+ RMSE: Là sai số bình phương trung bình (Root Mean Squared Error - RMSE) của mô hình trên tập huấn luyện và tập kiểm tra. Đây là một chỉ số đánh giá phổ biến khác, nó đo lường sự khác biệt trung bình giữa giá trị dự đoán và giá trị thực tế, sau khi bình phương sai số và tính toán căn bậc hai của giá trị này.

+ MSE: chỉ một trong những chỉ số được sử dụng phổ biến để đánh giá mức độ chính sát của mô hình dự đoán. MSE là tính giá trị trung bình của bình phương sai số giữa giá trị dự đoán và giá trị thực tế. Khi chỉ số MSE càng nhỏ thì mô hình dự đoán càng chính xác.

+ R2  score: Là chỉ số được dùng để đo lường mức độ phù hợp của mô hình dự đoán và dữ liệu thực tế. Nó cho biết tỉ lệ phần trăm phương sai của biến phụ thuộc.

A picture containing text, screenshot, software, font

Description automatically generated

A picture containing text, screenshot, font

Description automatically generated

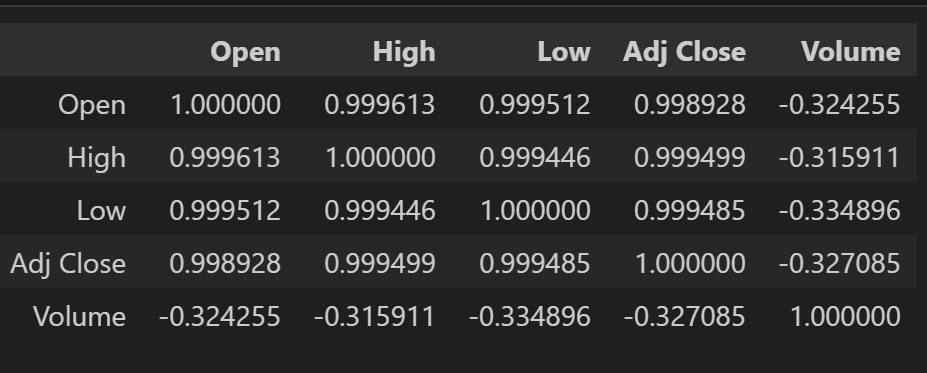
- Tính độ tương quan giữa biến phụ thuộc và các biến độc lập

+ Giá trị mở cửa (Open), giá cao nhất (High), giá thấp nhất (Low) và giá đóng cửa điều chỉnh (Adj Close) có mối tương quan rất cao với nhau. Các giá trị này có tương quan dương mạnh (gần 1), cho thấy sự tương đồng lớn giữa chúng. Điều này có ý nghĩa rằng khi một giá trị tăng/giảm, các giá trị còn lại cũng có xu hướng tăng/giảm tương tự.

+ Giá đóng cửa điều chỉnh (Adj Close) có mối tương quan cao đến mức đáng kể với giá mở cửa (Open), giá cao nhất (High) và giá thấp nhất (Low). Điều này cho thấy giá đóng cửa điều chỉnh có thể được dự đoán một cách chính xác dựa trên giá mở cửa, giá cao nhất và giá thấp nhất.

+ Khối lượng giao dịch (Volume) có mối tương quan tiêu cực với các giá trị khác (-0.324255 đến -0.334896). Điều này có nghĩa là khi giá trị khối lượng giao dịch tăng, giá trị của giá mở cửa, giá cao nhất, giá thấp nhất và giá đóng cửa điều chỉnh có xu hướng giảm.

Tóm lại, các biến giá cổ phiếu (Open, High, Low, Adj Close) có mối tương quan cao và có thể được sử dụng để dự đoán giá đóng cửa điều chỉnh (Adj Close). Tuy nhiên, khối lượng giao dịch (Volume) có tương quan tiêu cực với các biến giá, có thể ảnh hưởng đến dự đoán giá cổ phiếu.



* Biểu đồ thể hiện sự tương quan giữa các biến (Open, High, Low, Close): cho thấy mức nhìn tổng quan hơn về mức độ tương quan giữa các biến.

A picture containing text, screenshot, number

Description automatically generated

* Hiển thị 5 dòng dầu tiên giá trị thực tế của cổ phiếu Apple từ tập kiểm tra cho ta cái nhìn ban đầu về giá trị thực tế và giá trị dự đoán của cổ phiếu đó.

A picture containing text, screenshot, font, number

Description automatically generated

* Để định lượng tính hữu ích của tất cả các biến trong toàn bộ khu rừng ngẫu nhiên, chúng ta có thể xem xét tầm quan trọng tương đối của các biến.
* Các tầm quan trọng được trả về trong Skicit-learning thể hiện mức độ bao gồm một biến cụ thể cải thiện dự đoán.

A picture containing text, screenshot, diagram, rectangle

Description automatically generated  
Biểu đồ thể hiện các biến quan trọng trong mô hình

* Vẽ biểu đồ dự đoán và nhận xét:

*A picture containing text, screenshot, line, font

Description automatically generated*

Để dự báo giá cổ phiếu của Apple theo mô hình Random Forest, bộ dữ liệu thu thập được về các biến trong mô hình được chia thành 2 tập là tập huấn luyện (training dataset), chiếm 80%, được dùng để điều chỉnh các tham số của mô hình; và tập kiểm tra (testing dataset), chiếm 20%, được dùng để đánh giá mô hình. Tập mẫu được rút ra từ tập huấn luyện để xây dựng các cây quyết định. Mô hình được huấn luyện dựa trên tập huấn luyện (training dataset). Sau khi kết thúc quá trình huấn luyện, mô hình sẽ được đánh giá hiệu quả trên cơ sở tập kiểm tra (testing dataset) bằng cách so sánh giá trị dự báo và giá trị thực tế của biến phụ thuộc và các chỉ tiêu đánh giá sai số dự báo.

Kết quả dự báo theo mô hình Random Forest được thể hiện trong Hình . Theo đó, giá cổ phiếu được dự báo theo tập huấn luyện bám rất sát giá cổ phiếu thực tế. Mô hình dự báo được kiểm tra lại trên cơ sở tập kiểm tra. Tuy độ chính xác của các dự báo không bằng so với kết quả thu được từ tập huấn luyện, nhưng kết quả dự báo giá cổ phiếu từ tập kiểm tra cũng rất gần với giá giá cổ phiếu Apple thực tế.

Để đánh giá hiệu suất của mô hình Random Forest trong việc dự báo giá cổ phiếu, chúng tôi sử dụng các chỉ tiêu đánh giá sai số dự báo là sai số trung bình tuyệt đối (MAE) và căn bậc hai trung bình bình phương sai số dự báo (RMSE). Các sai số dự báo từ tập huấn luyện, MAE = 0.22 và RMSE =0.42. Trong khi đó, dự báo từ tập kiểm tra có MAE = 1.38 và RMSE =1.76, cao hơn so với các sai số dự báo tương ứng từ tập huấn luyện, nhưng vẫn ở mức thấp cho thấy mô hình có khả năng dự đoán khá chính xác.

1. **Kết luận**

Bài báo này vận dụng mô hình trí tuệ nhân tạo Random Forest để dự báo giá cổ phiếu thế giới dựa trên 7 nhân tố ảnh hưởng tới giá giá cổ phiếu Apple, bao gồm ngày giao dịch, giá mở cửa, giá đóng cửa, giá đóng cửa có điều chỉnh, giá cổ phiếu cao nhất trong ngày giao dịch, giá thấp nhất trong ngày giao dịch và khối lượng giao dịch của cổ phiếu Apple trong ngày. Kết quả nghiên cứu cho thấy mô hình Random Forest được đề xuất trong nghiên cứu này có thể dự báo tốt giá cổ phiếu của Apple. Kết quả dự báo từ tập huấn luyện và tập kiểm tra đều bám sát giá cổ phiếu thực tế trong giai đoạn nghiên cứu với sai số dự báo (MAE và RMSE) thấp.

**TÀI LIỆU THAM KHẢO**

1. Robert E. Schapire, 2001, Random Forest, Statistics Department, University of California, Berkeley, CA 94720
2. TS. Vũ Diệp Anh, 2022, Ứng dụng mô hình Random Forest để dự báo giá giá cổ phiếu Apple thế giới. Tạp Chí Công Thương, https://tapchicongthuong.vn/bai-viet/ung-dung-mo-hinh-random-forest-de-du-bao-gia-quang-sat-the-gioi-89938.htm