TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN PHÂN TÍCH XÁC SUẤT**

**VÀ GIẢI THUẬT NGẪU NHIÊN**

**Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500**

*Người hướng dẫn*: **T.S NGUYỄN CHÍ THIỆN**

*Người thực hiện*: **TRƯƠNG TRỌNG VINH - 186005039**

**LÊ ANH KHOA - 186005031**

Lớp **: 18600531**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2019**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN PHÂN TÍCH XÁC SUẤT**

**VÀ GIẢI THUẬT NGẪU NHIÊN**

**Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500**

Người hướng dẫn: **T.S NGUYỄN CHÍ THIỆN**

Người thực hiện: **TRƯƠNG TRỌNG VINH**

**LÊ ANH KHOA**

Lớp **:** **18600531**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2019**

LỜI CẢM ƠN

Trong thời gian làm nghiên cứu, chúng tôi đã nhận được nhiều sự giúp đỡ, đóng góp ý kiến và chỉ bảo nhiệt tình của thầy cô, bạn bè và anh chị em đồng nghiệp

Chúng tôi xin gửi lời cảm ơn chân thành đến T.S Nguyễn Chí Thiện, giảng viên  
khoa Công Nghệ Thông Tin - trường ĐH Tôn Đức Thắng người đã tận tình hướng dẫn chúng tôi trong suốt quá trình làm đề tài cuối khóa.

Chúng tôi cũng xin chân thành cảm ơn các thầy cô giáo trong trường ĐH Tôn  
Đức Thắng nói chung, các thầy cô trong khoa Công Nghệ Thông Tin nói riêng đã chỉ dạy và cung cấp những kiến thức cơ bản về Phân tích xác suất và giải thuật ngẫu nhiên số để chúng tôi có một nền tảng vững chắc thực hiện tốt đề tài cuối khóa này.

Cuối cùng, chúng tôi xin chân thành cảm ơn bạn bè và đồng nghiệp đã luôn tạo điều kiện, quan tâm, giúp đỡ, động viên chúng tôi trong suốt quá trình học tập và hoàn thành đề tài.

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Chúng tôi xin cam đoan đây là công trình nghiên cứu của riêng chúng tôi và được sự hướng dẫn khoa học của T.S Nguyễn Chí Thiện. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong luận văn còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào chúng tôi xin hoàn toàn chịu trách nhiệm về nội đề tài của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 16 tháng 6 năm 2019*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Trương Trọng Vinh*

*Lê Anh Khoa*

TÓM TẮT

Trong đề tài được giao lần này, chúng tôi đã hiện thực được các nội dung sau:

Nghiên cứu bài báo khoa học “Krauss, Christopher, Xuan Anh Do, and Nicolas Huck. “Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500.” European Journal of Operational Research (2017)”.

Xây dựng chương trình mô phỏng theo bài báo.

Hiểu được các thuật toán và kiến thức mà bài báo mang lại.

MỤC LỤC

[CHƯƠNG 1 – TÓM TẮT NỘI DUNG 4](#_Toc11961732)

[CHƯƠNG 2 – GIỚI THIỆU 5](#_Toc11961733)

[CHƯƠNG 3 – TỔNG QUAN TÀI LIỆU 7](#_Toc11961734)

[CHƯƠNG 4 – DỮ LIỆU VÀ PHẦN MỀM 8](#_Toc11961735)

[**4.1 Dữ liệu** 8](#_Toc11961736)

[**4.2 Phần mềm** 8](#_Toc11961737)

[CHƯƠNG 5 – PHƯƠNG PHÁP LUẬN 9](#_Toc11961738)

[**5.1 Các bước thực hiện** 9](#_Toc11961739)

[**5.2 Tạo ra bộ training và trading.** 9](#_Toc11961740)

[**5.3 Sinh đặc trưng:** 9](#_Toc11961741)

[**5.4 Traing model** 10](#_Toc11961742)

[**5.4.1 Mô hình học sâu** 10](#_Toc11961743)

[**5.4.2 Gradient-boosted trees** 11](#_Toc11961744)

[**5.4.3 Rừng ngẫu nhiên** 12](#_Toc11961745)

[**5.4.4 Tập hợp cân bằng.** 13](#_Toc11961746)

[**5.4.5 Dự báo, xếp hạng và giao dịch** 13](#_Toc11961747)

[CHƯƠNG 6 – KẾT QUẢ 15](#_Toc11961748)

[**6.1 Tổng quan:** 15](#_Toc11961749)

[**6.2 Phân tích** 15](#_Toc11961750)

[**6.2.1 Phân tích chung** 15](#_Toc11961751)

[**6.2.2 Lợi nhuận hằng ngày:** 16](#_Toc11961752)

[**6.2.3 Lợi nhuận hằng năm** 17](#_Toc11961753)

[**6.2.4 Tiếp xúc lợi nhuận với các nguồn rủi ro hệ thống phổ biến** 18](#_Toc11961754)

[CHƯƠNG 7 – KẾT LUẬN 22](#_Toc11961755)

**DANH MỤC VIẾT TẮT**

*DNN: Deep Neural Network*

*GBT: Gradient-boosted trees*

*RAF: Random Forest*

*S&P 500: Standard & Poor's 500 Stock Index*

**MỤC LỤC HÌNH ẢNH**

*Hình 1: Mô phỏng quá trình tạo ra bộ training và trading* 7

*Hình 2: Mô phỏng mô hình Deep Neural Network* 9

*Hình 3: Mô phỏng mô hình Gradient-boosted trees* 10

*Hình 4: Mô phỏng mô hình Random Forest* 11

*Hình 5: Biểu đồ phân tích chung toàn bài toán* 13

*Hình 6: Số liệu lợi nhuận hằng ngày.* 14

*Hình 7: Số liệu lợi nhuận hằng năm.* 15

*Hình 8: Số liệu tiếp xúc lợi nhuận và nguồn rủi ro phổ biến* 16

*Hình 9: Chỉ số giai đoạn phụ* 17

*Hình 10: Số liệu biến quan trọng.* 18

*Hình 11: Thông số mô hình.* 18

CHƯƠNG 1 – TÓM TẮT NỘI DUNG

Trong những năm gần đây, nghiên cứu về máy học mà cụ thể là phương pháp học sâu đã dần thay thế các phương pháp truyền thống trong việc ứng dụng vào nhiều ngành nghề khác nhau mà trong đó phân tích tài chính là một ví dụ. Trong phạm vi bài viết này, tác giả thực hiện việc phân tích hiệu quả của các thuật toán: học sâu (DNN), cây gradientboosted (GBT), rừng ngẫu nhiên (RAF) và sự kết hợp (ENS) trong việc thống kê số liệu. Mỗi mô hình đều được ứng dụng trong việc phân tích lợi nhuận sau cùng của các cổ phiếu trong chỉ số S&P 500 sau khi loại bỏ các điều kiện không cần thiết. Từ năm 1992 đến 2015, các tính hiệu giao dịch hằng ngày được tạo ra dựa trên một báo cáo xác xuất một số cổ phiếu vượt trội hơn thị trường chung. Và xác suất K cao được chuyển đổi thành long-position và K thấp thành các short-position.

Kết quả thực nghiệm cho thấh đầy hứa hẹn, các thuật toán được áp dụng tạo ra lợi nhuận vượt mẫu 0,45% mỗi ngày với K=10. Bất kể việc thực tế là lợi nhuận đang giảm trong vài năm gần đây, những phát hiện của tác giả đã đặt ra một thách thức đầy quan trọng đối với hình thức bán ra dựa trên hiệu quả của thị trường.

CHƯƠNG 2 – GIỚI THIỆU

Thống kê trao đổi chứng khoán là một thuật ngữ bao quát được triển khai trong các giao dịch dự phòng hay các giao dịch độc quyền. Nó bao gồm các tính năng sau: Tín hiệu giao dịch, trái với các điều kiện cơ bản, không có thử nghiệm với thị trường và tạo ra những lợi nhuận vượt mức thống kê. Điều này liên quan đến số lượng chứng khoán, thời gian nắm giữ ngắn và tính toán, giao dịch đáng kể, cơ sở hạ tầng CNTT. Các mô hình cơ bản có tính độc quyền cao và điều đó hạn chế quyền truy cập của công chúng hay các nhà khoa học. Ngược lại vấn đề đó, nghiên cứu tài chính cổ điển tập trung vào giá trị bất ổn của thị trường với việc giải thích rõ ràng.

Phạm vi trong bài nghiến cứu này, tác giả đã phát triển một chiến lược thống kê chênh lệch ngắn hạn cho các thành phần S&P 500. Với mục đích đó, tác giã đã lấy cảm hứng từ các phương pháp như học sâu (DNN) là một loại mô hình được số hóa đi qua nhiều lớp ẩn và cho ra tỉ lệ lỗi là thấp hơn 1%.

Năm 1997, Deep Blue đã đánh bại nhà vô địch cờ vây thế giới Garry Kasparov, những trò chơi như vậy có thể được giải quyết bẳng cách tính toán đệ quy hàm giá trị tối ưu trong cây tìm kiếm và trong năm 2016 Alpha Go ứng dụng hệ thống thần kinh sâu và cây Monte Carlo đã đánh bại các nhà vô định cờ vây ở Châu Âu vào các năm 2013, 2014 và 2015.

Thứ hai, các giả ứng dụng cây gradientboosted. Về cơ bản nó kết hợp nhiều học yếu thành học mạnh. Thứ ba, tác giả ứng dụng thuật toán rừng ngẫu nhiên như là một sự điều chỉnh đáng kể trong các phương pháp trên bằng cách xây dựng các cây không tương quan với nhau. Cuối cùng là kết hợp ba phương pháp đó lại thành một nhóm đơn giản.

Tác giải đào tạo ra các mô hình này với lợi nhuận cuối cùng của các thành phần trong chỉ số S&P 500 và dự báo xác suất cho các cổ phiếu vượt trội. Các cổ phiếu có K hàng đầu được mua và các cổ phiếu cho K giảm thì được bán. Đối với cả ba mô hình với K=10, tác giả nhận thấy lợi nhuận trung bình là 0.45% trước mỗi phiên giao dịch với các thông số: học sâu là 0.33%, cây gradientboosted là 0.37% và rừng ngẫu nhiên là 0.43%. Do tần suất giao dịch lớn nên lợi nhuận sẽ còn 0.25% sau khi trừ đi các chi phí giao dịch. Ngoài ra tác giải còn nhận thấy có những đột biến tích cực trong lúc thị trường đang hỗn loạn nhưu bóng bóng dotcom hay khủng hoàng tài chính toàn cầu. Từ 2001 với sự phổ biến của các phương pháp máy học và sức mạnh tính toán ngày càng tăng, tác giả nhận thấy lôi nhuận giảm dần cho thấy hiệu trường đã đi sát hơn với các dự đoán được đưa ra trước đó.

CHƯƠNG 3 – TỔNG QUAN TÀI LIỆU

Tài liệu được sử dụng bao gồm Huck (2009,2010), Takeuchi và Lee (2013); Moritz và Zimmermann (2014); Dixon và cộng sự. (2015), cung cấp các ứng dụng ban đầu của các kỹ thuật máy học để phân xử thống kê.

Huck (2009) đã phát triển một thuật toán thóng kê chênh lệch dựa trên hệ thần kinh Elman và ElectroRE III, một phương pháp đa tiêu chí. Phương pháp của ông bao gồm dự báo, vượt trội và giao dịch. Trong bước dự báo, Huck (2009) sử dụng các mạng thần kinh để tạo dự báo hoàn vốn trước một tuần. ElectroRE III được sử dụng để tạo ra sự vượt trội của tất cả các cổ phiếu dựa trên ma trận đầu vào này. Với hiệu suất tương đối các cổ phiếu bị định giá thấp sẽ vượt lên ở các cổ phiếu hàng đầu và được định giá cao ở cuối bảng xếp hạng đó.

Takeuchi và Lee (2013) mạng lưới thần kinh sâu được sử dụng làm phân loại để tính xác suất cho mỗi cổ phiếu vượt trội so với lợi nhuận trung bình của tất cả các cổ phiếu trong tháng giữ t + 1. Không gian tính năng được tạo như sau: Cứ mỗi tháng t, các tác giả xây dựng chuỗi thời gian trả lại tích lũy được tiêu chuẩn hóa cho 12 tháng t - 2 đến t - 13 và 20 ngày qua tương ứng với tháng t. Cùng với một biến giả biểu thị nếu thời gian nắm giữ của tháng t + 1 tương ứng với tháng 1, tổng cộng 33 dự đoán được tạo. Chúng đạt độ chính xác phân loại đáng kể là 73 phần trăm, mặc dù không xem xét các hiệu ứng vi cấu trúc - vốn là tinh túy trong ánh sáng của dữ liệu tần số cao.

Moritz và Zimmermann (2014) triển khai các khu rừng ngẫu nhiên trên dữ liệu CRSP của Hoa Kỳ từ năm 1968 đến 2012 để phát triển một chiến lược giao dịch dựa trên danh mục đầu tư có điều kiện. Một khu rừng ngẫu nhiên được đào tạo để dự đoán lợi nhuận cho mỗi cổ phiếu trong 12 tháng sau khi hình thành danh mục đầu tư.

Krauss (2015) cung cấp một đánh giá gần đây về hơn 90 chiến lược giao dịch chênh lệch giá theo cặp thống kê, tập trung vào các sai lầm tương đối giữa hai và nhiều chứng khoán. Atsalakis và Valavanis (2009) khảo sát hơn 100 bài viết sử dụng các kỹ thuật máy học để dự báo thị trường chứng khoán.

**CHƯƠNG 4 – DỮ LIỆU VÀ PHẦN MỀM**

**4.1 Dữ liệu**

Theo kinh nghiệm, tác giả chọn S&P 500, vì S&P 500 là tổ hợp của 500 công ty lớn nhất thế giới chiếm 80% vốn hóa của thụ trường. Do đó các tập hợp thanh khoản có tác dụng như một thử nghiệm đúng nghĩa cho thị trường. Đầu tiên tác giả thu thập danh sách tạo ra S&P 500 từ Thomson Reuters Datastream từ 12/1989 đến 9/2015 và tạo chúng thành một ma trận nhị phân cho biết liệu cổ phiếu có phải là thành phần của chỉ số trong tháng tiếp theo hay không. Thứ hai đối với các cổ phiếu đã từng là một thành phần của S&P 500 thì tác giả thu thấp các chỉ số tổng lợi nhuận hàng ngày từ 1/1990 đến 10/2015 giá cổ phiếu này phản ánh giá cổ tức và tài khoản cho tất cả các hành động của công ty và chia tách cổ phiếu. Ngoài ra các ngày nghỉ lễ cũng được loại bỏ.

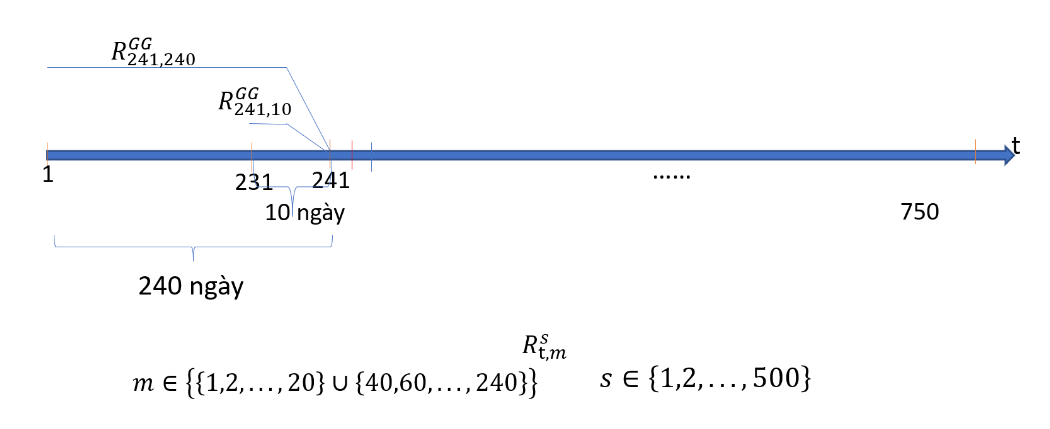
**4.2 Phần mềm**

Tiền xử lý dữ liệu được xử lý bằng ngôn ngữ R. Để tập hợp chuỗi thời gian tác giả sử dụng các gói xts của Ryan và Ulrich(2014) và TTR của Ulrich(2013). Đánh giá hiệu suất bẳng gói PerformanceAnalytics của Peterson và Carl(2014). Ba thuật toán được sử dụng thông qua H2O trên nền Java để có thể triển khai nhanh và dễ dàng mở rộng, nền tảng này đang được triển khai trong hơn 2000 công ty. Một phần của R và H2O được triển khai thông qua Windows PowerShell.

**CHƯƠNG 5 – PHƯƠNG PHÁP LUẬN**

**5.1 Các bước thực hiện**

* Chia toàn bộ dữ liệu trong bộ training và trading không chồng lên nhau.
* Bộ training dùng để train và bộ trading dùng cho ứng dụng ngoài mẫu.
* Với mội cặp training và trading tác giả tạo không gian để dự đoán.
* Sử dụng các thuật toán DNN, GBT, RAF trên mỗi bộ.
* Sử dụng các kết quả trên với các bộ trading tương ứng để đưa ra dự đoán.

**5.2 Tạo ra bộ training và trading.**

*Hình 1: Mô phỏng quá trình tạo ra bộ training và trading*

Tác giả tạo ra các cửa sổ training và 750 ngày (3 năm) và cửa sổ trading là 250 ngày (2 năm) và các cửa sổ này di chuyển theo dạng trượt trừ 1990 đến 2015. Một tâm dữ liệu huấn luyện bao gồm 500 \* 750 = 375000 và một tập giao dịch là 125000 mẫu.

**5.3 Sinh đặc trưng:**

Đối với mỗi bộ giao dịch đào tạo, chúng tôi tạo không gian tính năng (đầu vào) và biến trả lời (đầu ra) như sau:

Trong đó:

: biểu thị lợi nhuận của cổ phiếu S trong khoảng thời gian m

biểu thị quá trình biến đổi về giá của chứng khoáng s với 1<=s<=500

Tổng cộng có 31 tính năng tương ứng một năm giao dịch khoảng 240 ngày.

**5.4 Traing model**

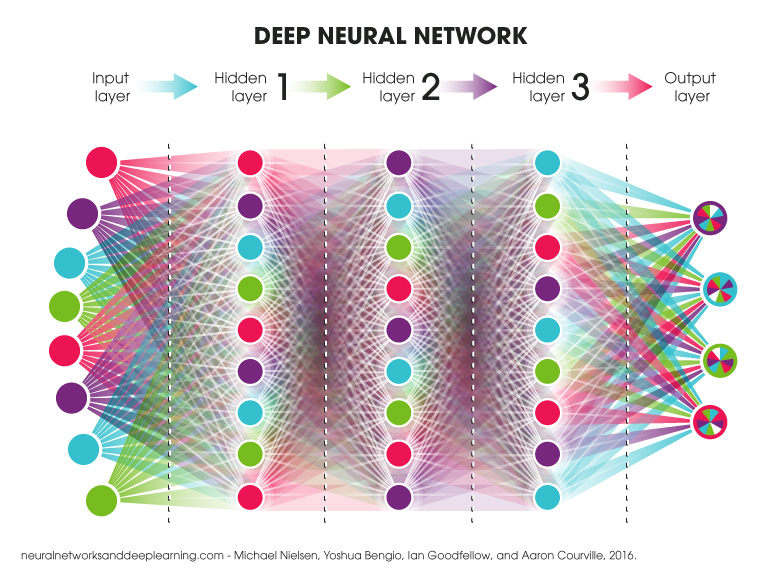
**5.4.1 Mô hình học sâu**

Theo Candel thì mô hình thần kinh sâu bao gồm một lớp đầu vào, một hoặc nhiều lớp ẩn và một lớp đầu ra, tạo thành cấu trúc liên kết của mạng. Lớp đầu vào khớp với không gian đặc trưng, ​​do đó có nhiều nơ ron đầu vào như các yếu tố dự đoán. Lớp đầu ra là lớp phân loại hoặc hồi quy để phù hợp với không gian đầu ra. Trong kiến ​​trúc feedforward cổ điển, mỗi nơ-ron ở lớp trước l được kết nối đầy đủ với tất cả các nơ-ron ở lớp tiếp theo l + 1 thông qua các cạnh được định hướng, mỗi cạnh biểu thị một trọng lượng nhất định. Ngoài ra, mỗi nơ-ron trong một lớp không đầu ra của mạng có một đơn vị sai lệch, đóng vai trò là ngưỡng kích hoạt của nó. Như vậy, mỗi nơ-ron nhận được tổ hợp trọng số α của đầu ra nl của các nơ-ron ở lớp trước làm đầu vào

Với wi biểu thị trọng lượng của đầu ra xi và b sai lệch. Tổ hợp trọng số α của (2) được biến đổi thông qua một số chức năng kích hoạt f, để tín hiệu đầu ra f (α) được chuyển đến các nơ ron trong lớp l + 1.

f(α1, α2) = max(α1, α2)

nhận đầu vào từ hai kênh riêng biệt với trọng số và độ lệch riêng. Sự lựa chọn của chúng tôi được thúc đẩy bởi thực tế là kích hoạt maxout có hiệu quả đặc biệt với trường hợp dropout

Đối với toàn bộ mạng, đặt W là tập hợp với biểu thị ma trận trọng số kết nối các lớp l và l + 1 cho một mạng gồm các lớp L. Tương tự, gọi B là tập

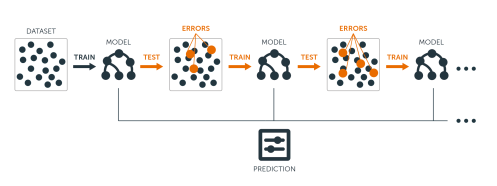
*Hình 2: Mô phỏng mô hình Deep Neural Network*

hợp với biểu thị vectơ cột của độ lệch cho lớp l. Bộ sưu tập W và B xác định đầy đủ đầu ra của toàn bộ DNN. Việc học được thực hiện bằng cách điều chỉnh các trọng số này để giảm thiểu lỗi trên dữ liệu huấn luyện. Cụ thể, mục tiêu là giảm thiểu một số hàm mất L (W, B | j) cho mỗi ví dụ đào tạo j. Vì chúng ta đang xử lý một vấn đề phân loại, nên hàm mất là entropy chéo

với y đại diện cho các đơn vị đầu ra và O lớp đầu ra.

**5.4.2 Gradient-boosted trees**

Gradient-boosted trees được giới thiệu trong một bài báo cáo chuyên đề của Schapire, mô tả một phương pháp kết hợp các thuật toán yếu thành các thuật toán đạt độ chính xác cao tùy ý.

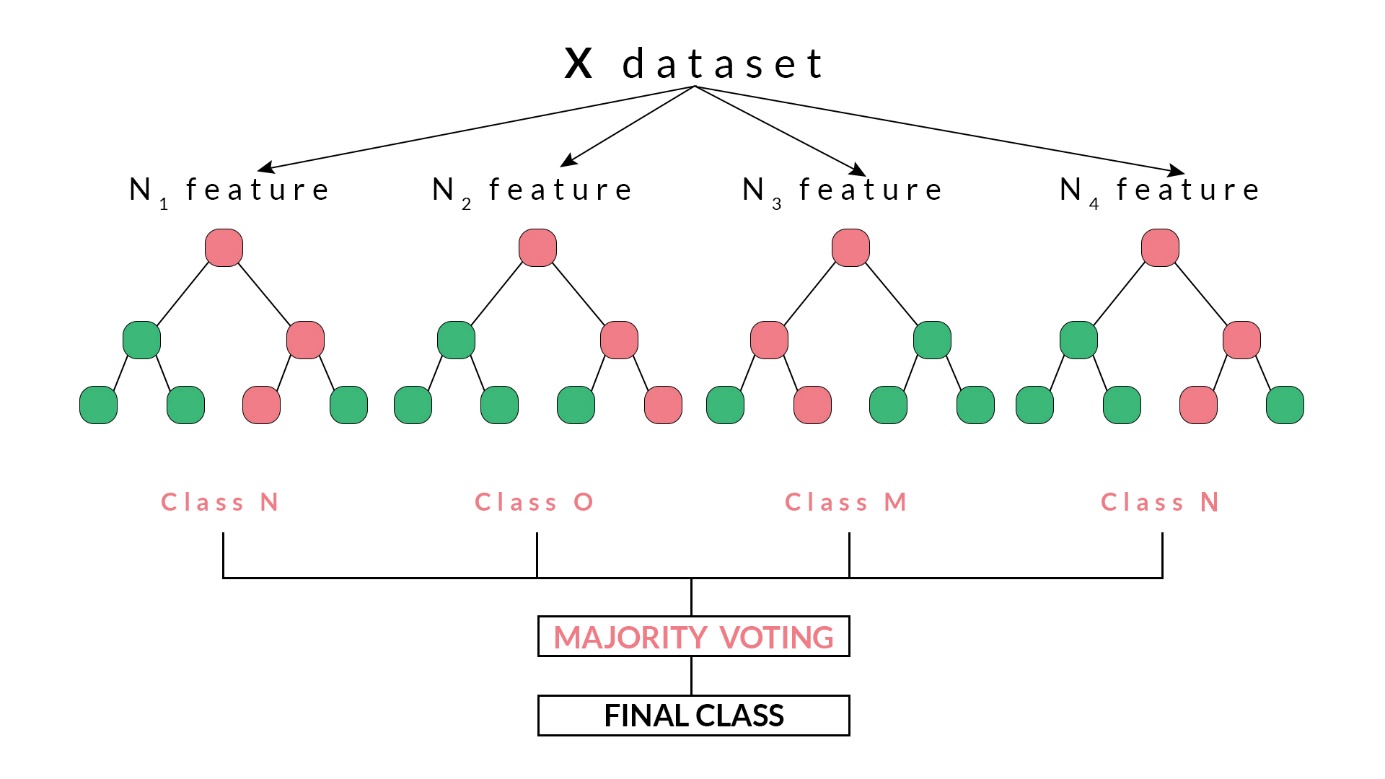


*Hình 3: Mô phỏng mô hình Gradient-boosted trees*

Ban đầu các vấn đề sẽ được phân loại và tăng cường hoạt động bằng việc áp dụng các thuật toán yếu với trọng số của dữ liệu được đào tạo. Sau mỗi lần lặp tăng cường, các ví dụ phân loại sai có trọng số tăng và các ví dụ được phân loại chính xác trọng lượng của chúng giảm. Do đó, mỗi phân loại kế tiếp tập trung vào các ví dụ khó phân loại trong các bước trước. Sau một số lần lặp MGBT, các dự đoán của một loạt các phân loại yếu được kết hợp bởi một phiếu bầu đa số có trọng số thành một dự đoán cuối cùng. Tăng cường độ dốc ngẫu nhiên là một biến thể được giới thiệu bởi Friedman trong đó các tác giả lấy mẫu một tập dữ liệu huấn luyện theo mỗi lần lặp để phù hợp với người học cơ sở. Các tác giả sử dụng cách tiếp cận hơi khác với truyền thống đó là chọn ngẫu nhiên các tính năng mGBT từ các tính năng p sau mỗi lần phân tách. Quy trình này làm tăng hiệu quả tính toán, giúp tăng hiệu suất và giải mã cây.

**5.4.3 Rừng ngẫu nhiên**

Trong thuật toán rừng ngẫu nhiên, các tác giả đã tạo ra các cây để diễn tả các dữ liệu, mỗi cây đều có tính lỗi của cây trước đó để xây dựng một tập hợp mạnh mẽ của các thuật toán. Các rừng ngẫu nhiên này sẽ bị phân ra và xây dựng trên nhiều mẫu dữ liệu khác nhau. Thuật toán để phát triển rừng ngẫu nhiên là tương đối đơn giản. Với mỗi cây BRAF trong rừng ngẫu nhiên, đầu tiw6n chúng ta sẽ tập hợp các con ngẫu nhiên từ dữ liệu ban đầu. Sau đó phát triển một cây quyết định đã sửa đỗi cho mẫu này và chọn các tính năng mRAF một cách ngẫu nhiên từ các tính năng p sau mỗi lần phân tách. Cây được tạo ra với độ sâu tối đa của JRAF.



*Hình 4: Mô phỏng mô hình Random Forest*

Đầu ra cuối cùng là các cây rừng ngẫu nhiên BRAF và việc phân loại sẽ được đưa ra dựa trên việc bỏ phiếu đa số. Việc lấy mẫu con làm giảm số phương sai của cây và lựa chọn ngẫu nhiên giúp phân rã chúng. Tác giả có ba tham số điều chỉnh cây BRAF, JRAF có độ sâu tối đa của số lương để chọn ngẫu nhiên mRAF. BRAF được chọn tối đa 1000 cây và JRAF được chọn ở mức 20 cây.

**5.4.4 Tập hợp cân bằng.**

Trong tập hợp cân bằng, tác giả đã lấy trung bình của ba mô hình đã sử dụng trước đó như sau:

Bàng cách sử dụng cách tính trung bình này tác giả đã giảm được rủi ro phân loại sai. Thứ hai là do cả ba phương pháp đều sử dụng phép thống kê ngẫu nhiên nên khi sử dụng phương pháp tính trung bình có thể đưa đến một kết quả xấp xỉ tốt hơn và từng mô hình một.

**5.4.5 Dự báo, xếp hạng và giao dịch**

Trong mỗi giai đoạn T+1 các tác giả dự báo cho mỗi cổ phiếu sẽ vượt trội hơn so với mặt bằng trung bình của mỗi cổ phiếu. Với mỗi ML ∈ {DNN, GBT, RAF, ENS} và s ∈ {1, . . . , n} sẽ sắp xếp tất cả các cổ phiếu theo thứ tự giảm dần. Ở đây các tác giả tìm thấy các cổ phiếu bị định giá thấp nhất theo mô hình học tập tương ứng và ở dưới cùng các cổ phiếu được định giá cao nhất với các xác xuất thấp nhất để vượt qua trung bình mặt cắt ngang trong giai đoạn t + 1.

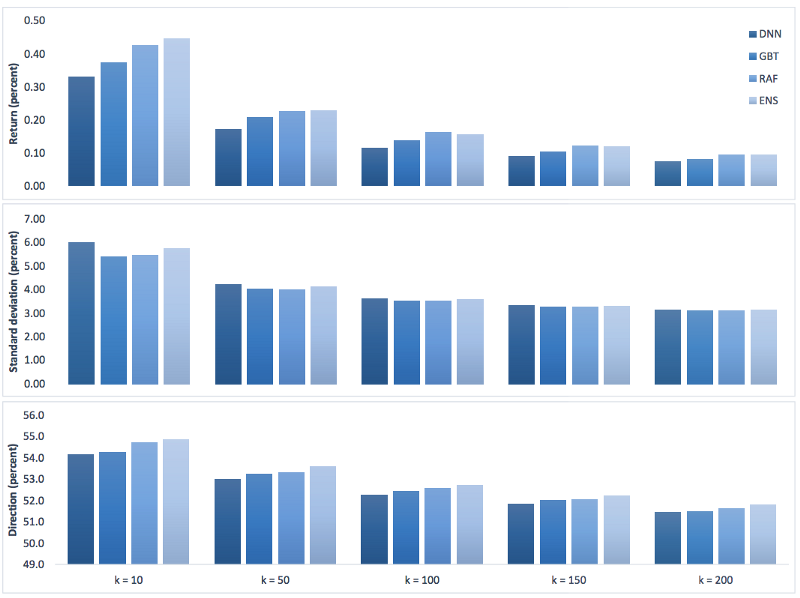
**CHƯƠNG 6 – KẾT QUẢ**

**6.1 Tổng quan:**

Với các tài liệu đã nghiên cứu, nhóm tác giả đã đóng góp 3 vấn đề như sau:

* Đây là nghiên cứu duy nhất trong việc triển khai ba kỹ thuật học máy hiện đại và tập hợp chúng trên một dữ liệu chứng khoáng lớn, do đó nhóm tác giả có thể so sánh hiệu suất của các phương pháp với nhau
* Nhóm tác giả cung cấp một đánh giá hiệu suất toàn diện, theo các tiêu chuẩn hiện hành trong tài liệu tài chính. Nó sẽ tiết lộ rằng lợi nhuận của các tập đoàn chỉ là một phần và các rủi ro có hệ thống, các chi phí giao dịch sẽ xấu đi theo thời gian – đó là nhờ sự xuất hiện của các phương pháp học máy và những tiến bộ trong sức mạnh tính toán. Tuy nhiên điều nhận được lớn nhất là các chuyển biến tích cực mạnh mẽ vẫn có thể thấy được trong những thời điểm thị trường có sự biến động cao.
* Thứ ba là nhóm tập trung vào việc đầu tư hằng ngày hơn là đầu tư hằng tháng giúp khai thác lợi nhuận trong ngắn hạn.

**6.2 Phân tích**

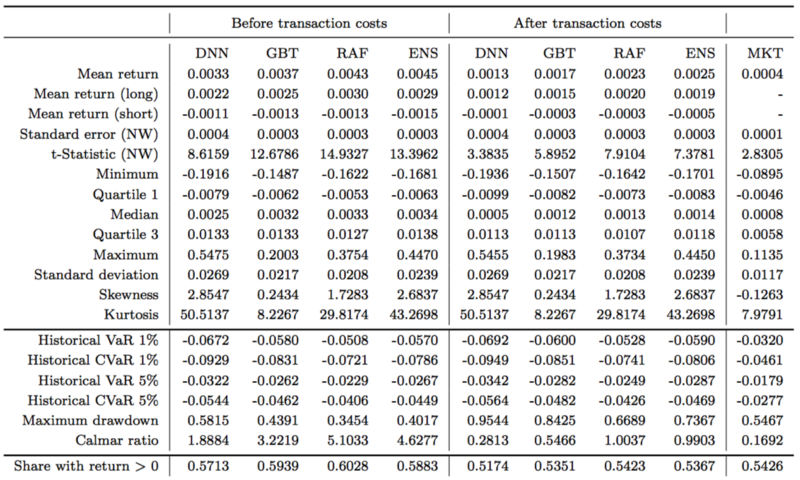
**6.2.1 Phân tích chung**

*Hình 5: Biểu đồ phân tích chung toàn bài toán*

Hình trên cho thấy số liệu hiệu suất hàng ngày cho các danh mục đầu tư ngắn có kích thước khác nhau: Lợi nhuận trung bình, độ lệch chuẩn và độ chính xác định hướng từ tháng 12 năm 1992 đến tháng 10 năm 2015. Danh mục đầu tư bao gồm các cổ phiếu k hàng đầu, với k ∈ {10, 50, 100, 150, 200}

Với k = 10 (với 10 vị trí dài và 10 vị trí ngắn). Kết quả quan sát thấy rằng quần thể tạo ra lợi nhuận 0,45% mỗi ngày, tiếp theo là rừng ngẫu nhiên với 0,43%, cây được tăng cường độ dốc với 0,37% và mạng lưới thần kinh sâu với 0,33% mỗi ngày.

Những kết quả này cho thấy việc tăng k, tức là, bao gồm các cổ phiếu có độ không chắc chắn cao hơn, dẫn đến giảm lợi nhuận và độ chính xác định hướng.

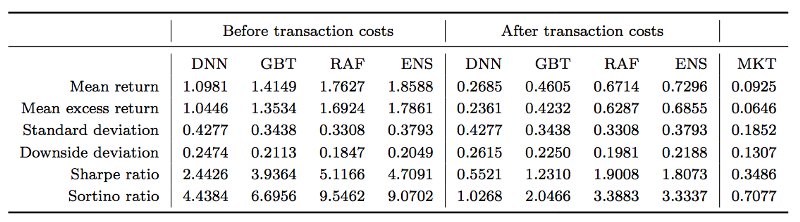
**6.2.2 Lợi nhuận hằng ngày:**

*Hình 6: Số liệu lợi nhuận hằng ngày.*

Đặc điểm hoàn trả hàng ngày của danh mục k = 10, trước và sau chi phí giao dịch cho DNN, GBT, RAF, ENS với thị trường chung (MKT) từ tháng 12 năm 1992 đến tháng 10 năm 2015. NW biểu thị các lỗi tiêu chuẩn Newey-West với sửa lỗi một độ trễ.

Về mặt này, RAF thể hiện rủi ro thấp nhất và DNN cao nhất. So với các chiến lược khác, chẳng hạn như giao dịch cặp cổ điển, rủi ro đuôi là đáng kể.

Giao dịch cặp dựa trên khoảng cách là một chiến lược cân bằng, có xu hướng xây dựng các cặp thể hiện sự biến động thấp. Tuy nhiên, rủi ro thấp hơn có lợi nhuận thấp (giao dịch cặp cổ điển chỉ đạt được lợi nhuận vượt quá hàng năm là 11%), do đó lợi nhuận thấp hơn đi cùng với rủi ro đuôi thấp hơn. Kết quả này áp dụng tương tự cho giá trị có điều kiện có rủi ro (CVaR).

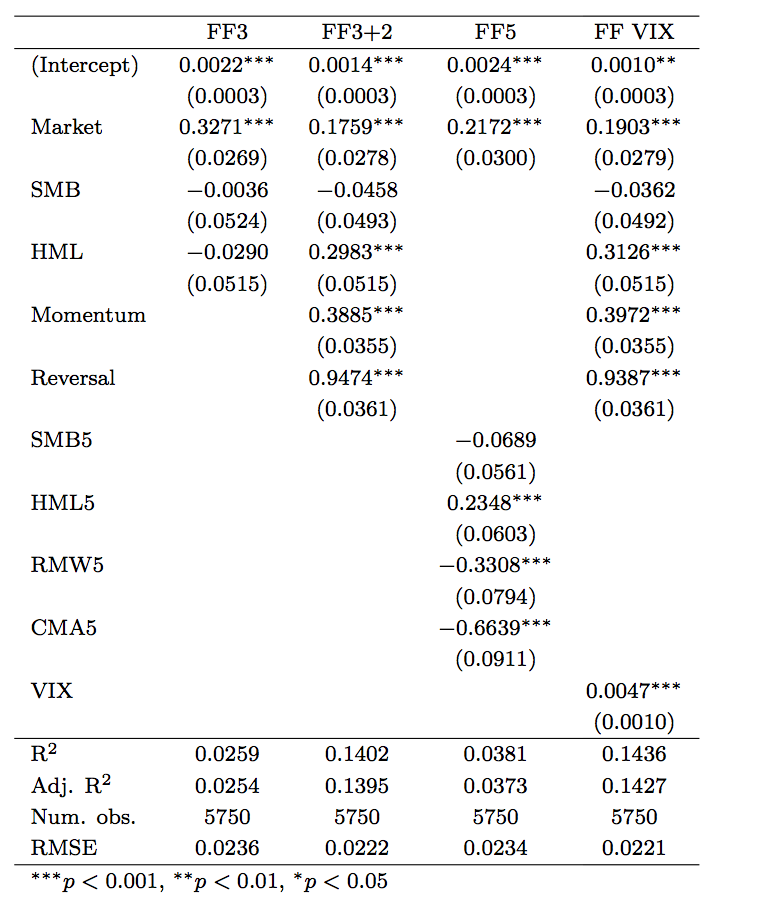
**6.2.3 Lợi nhuận hằng năm**

*Hình 7: Số liệu lợi nhuận hằng năm.*

Sau chi phí giao dịch, lợi nhuận hàng năm lên tới 73% cho toàn bộ, so với 67% đối với rừng ngẫu nhiên, 46% đối với cây tăng cường độ dốc và 27% đối với mạng lưới thần kinh sâu. Tất cả các chiến lược phần lớn vượt trội so với thị trường chung với lợi nhuận trung bình 9% p.a.

Lợi nhuận hàng năm và các biện pháp rủi ro của danh mục đầu tư k = 10, trước và sau chi phí giao dịch cho DNN, GBT, RAF, ENS so với thị trường chung (MKT) từ tháng 12 năm 1992 đến tháng 10 năm 2015.

Chúng tôi thấy rằng độ lệch nhược điểm nằm trong khoảng từ 0,20 (RAF) đến 0,26 (DNN), với giá trị 0,22 cho toàn bộ - khoảng 1,7 lần mức của thị trường chung. Do đó, độ lệch nhược điểm ít được thể hiện cho các chiến lược học máy so với các chiến lược của thị trường chung, dẫn đến tỷ lệ Sortino thuận lợi cho các mục tiêu mô hình học máy hoặc ít nhất là bao gồm cả sự phát triển này.

**6.2.4 Tiếp xúc lợi nhuận với các nguồn rủi ro hệ thống phổ biến**

*Hình 8: Số liệu tiếp xúc lợi nhuận và nguồn rủi ro phổ biến*

Chiến lược tập hợp với k = 10: Tiếp xúc với các nguồn rủi ro có hệ thống sau chi phí giao dịch cho DNN, GBT, RAF, ENS từ tháng 12 năm 1992 đến tháng 10 năm 2015. Lỗi tiêu chuẩn được mô tả trong ngoặc đơn.

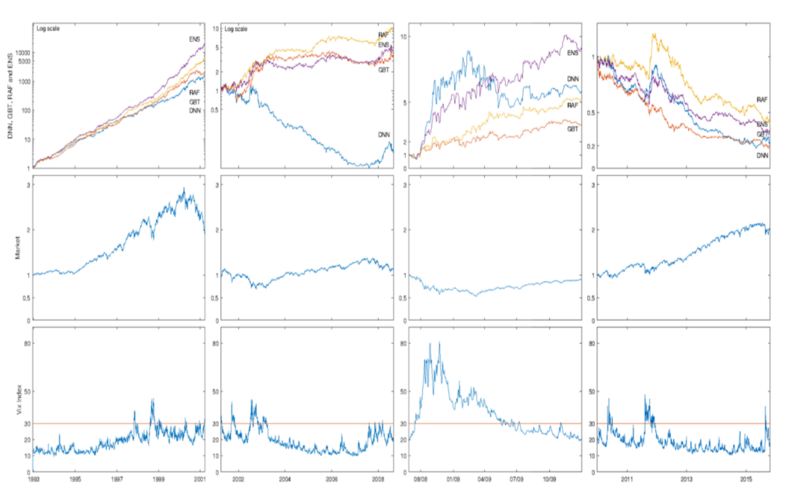
Họ kết luận rằng chiến lược tập hợp tạo ra các bảng chữ cái hàng ngày có ý nghĩa thống kê và kinh tế trong khoảng từ 0,14 đến 0,24% - tùy thuộc vào mô hình nhân tố được sử dụng.

6.2.5 Phân tích giai đoạn phụ

12/92 đến 03/01 tương ứng với thời kỳ vượt trội mạnh mẽ và nhất quán, trước khi phát minh và tuyên truyền các thuật toán học máy được sử dụng trong bài báo này.

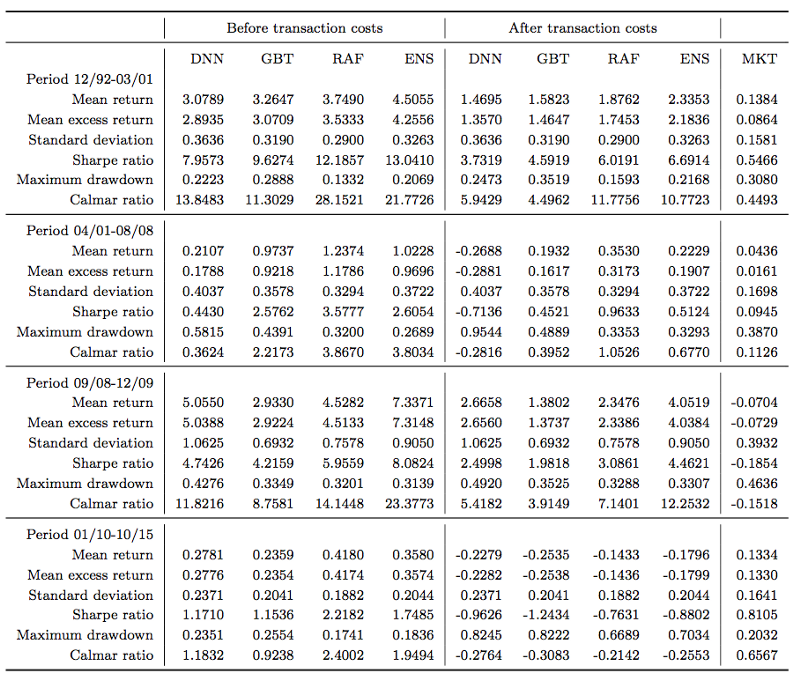
04/01 đến 08/08 và tương ứng với thời gian sửa đổi

09/08 đến 12/09 và tương ứng với cuộc khủng hoảng tài chính toàn cầu.

01/01 đến 10/15 và tương ứng với thời gian phát hiện.

*Hình 9: Chỉ số giai đoạn phụ*

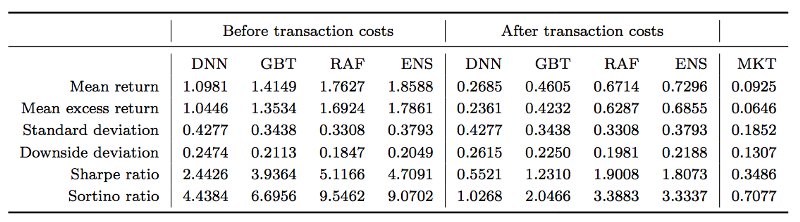
Hồ sơ phụ của danh mục k = 10, sau chi phí giao dịch cho DNN, GBT, RAF, ENS so với thị trường chung (MKT) và chỉ số VIX từ tháng 12 năm 1992 đến tháng 10 năm 2015.

6.2.6 Biến quan trọng

*Hình 10: Số liệu biến quan trọng.*

Đặc điểm lợi nhuận hàng năm cho mỗi giai đoạn đối với DNN, GBT, RAF,

6.2.7 Thông số mô hình

Bảng A: Thiết kế cấu hình đường cơ sở so với tham số hóa thấp hơn (phương án 1) và tham số hóa cao hơn (phương án 2). DNN cũng được so sánh với một mạng nơ-ron tiêu chuẩn với một lớp ẩn với 31 nơ-ron ẩn, không có chức năng kích hoạt thường xuyên và chức năng kích hoạt tanh (thay thế 3).

*Hình 11: Thông số mô hình.*

Bảng B: Lợi nhuận trung bình mỗi ngày cho danh mục k = 10 từ tháng 12 năm 1992 đến tháng 10 năm 2015 trước khi chi phí giao dịch.

CHƯƠNG 7 – KẾT LUẬN

Nhóm tác giả đã phát triển hệ thống thống kê dựa trên các mạng lưới DNN, cây gradient, rừng ngẫu nhiên từ tập hợp đơn giản đến triển khai trên các thành phần của chỉ số S&P 500 từ tháng 12 năm 1992 đến tháng 10 năm 2015. Nhóm tác giả có những đóng góp sau đây: Thuật toán rừng ngẫu nhiên có tính vượt trội hơn thuật toán cây gradient và phương pháp thần kinh sâu và thuật toán rừng ngẫu nhiên được chọn như thuật toán đặc trưng cho bài toán trong bài viết này. Thứ hai là kết hợp cả ba thuật toán này lại với nhau sẽ trở thành một tập hợp đơn giản và có kết quả tốt hơn so với sử dụng các thuật toán riêng lẻ. Trong trường hợp người dùng không biết chọn thuật toán nào thì việc kết hợp cả ba thuật toán này lại sẽ là một lựa chọn đúng đắn và mang tính đa dạng.

Đóng góp thứ hai đó là việc đánh giá hiệu suất của các chiến lược giao dịch. Đối với các giao dịch có k = 10, lợi nhuận dựa trên thống kê sẽ là 0.25% mỗi ngày hoặc 73% mỗi năm sau khi trừ đi các chi phí giao dịch. Có những tiềm năg để giảm độ lệch chuẩn và rủi ro sau cùng, ví dụ như tối ưu hóa danh mục đầu tư CvaR. Tiếp theo là những rủi ro có hệ thống nhưng sau khi trừ các rủi ro đó qua một bên thì lợi nhuận vẫn nằm trong 0.14% đến 0.24% mỗi ngày tùy thuộc vào mô hình người sử dụng. Tuy nhiên có những lúc lợi nhuận ngày càng ít đi do sự phát triển của máy học và các phương pháp tính toán kể từ năm 2001. Tuy nhiên khi có những bất ổn với thị trường như cuộc khủng hoảng nợ tài chính hoặc khủng hoảng Châu Âu thì thuật toán vẫn cho thấy tiềm năng của cổ phiếu trong lúc đó.

Đóng góp thứ ba là dựa trên tầng suất giao dịch hằng ngày cho ta thấy lợi nhuận tương gần đây nhất tương ứng với năm ngày giao dịch trước đó có giá trị cơ sở cao nhất. Sau đó, lợi nhuận tương ứng với 12 tháng giao dịch tiếp theo. Trong khi các bất thường của thị trường vốn tập trung vào dữ liệu hằng tháng thì trong nghiên cứu các tác giả đã chỉ ra rẳng các mô hình lợi nhuận có thể dựa trên lợi nhuận hằng ngày.

Nhìn chung các tác giả đã góp phần thu hẹp khoảng cách giữa học thuật và tài chính chuyên nghiệp. Lợi nhuận cao liên tục trước năm 2001 có thể có khả năng được hiểu là một chiến lược chênh lệch thống kê được hỗ trợ bởi máy học thành công cho đến những ngày đầu của thiên niên kỷ mới. Và sau đó là tập trung vào việc đầu cơ cổ phiếu do sự phát triển của máy học, không có sự rõ ràng nào về thông tin phân tích ở các công ty và không có sự minh bạch trong các giao dịch độc quyền hoặc trong các quỹ dự phòng. Tuy nhiên hướng phân tích của tác giả vẫn cho ra các kết quả tốt như Renaissance Technologies, một trong những quỹ phòng hộ định lượng thành công nhất, báo cáo lợi nhuận trung bình hàng năm là 71,8% từ năm 1994 đến giữa năm 2014 trước phí và tăng vọt 98,2% vào năm 2008, rõ ràng các tác giả đã thể hiện được hiệu suất mạnh mẹ và những thời điểm tăng đột biến trong lúc thị trường đang hỗn loạn

TÀI LIỆU THAM KHẢO

**Tiếng Anh**

1. Bài báo : Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500