Các yếu tố ảnh hưởng đến sai số trong dự báo tỷ suất sinh lời của cổ phiếu đơn lẻ: Ứng dụng học máy với Spark MLlib

Factor affecting the error in individual stock's return forecasting: Appling machine learning with Spark MLlib

Bùi Thành Khoa[1,2,4*], Trần Trọng Huỳnh[3], Thái Duy Tùng[4], Nguyễn Ngọc Dụng[5], Nguyễn Vũ Đức[4]

[1]Trường Đại học Công nghệ Thông tin, Thành phố Hồ Chí Minh, Việt Nam
[2]Trường Đại học Quốc gia Việt Nam, Thành phố Hồ Chí Minh, Việt Nam
[3]Trường Đại học FPT, Hà Nội, Việt Nam
[4]Trường Đại học Công nghiệp Thành phố Hồ Chí Minh, Việt Nam
[5]Trường Đại học Kinh tế Thành phố Hồ Chí Minh, Việt Nam
[*]Tác giả liên hệ, Email: buithanhkhoa@iuh.edu.vn; 19522611@gm.uit.edu.vn

THÔNG TIN

DOI:10.46223/HCMCOUJS.tech.vi.17.1.2245.2022

Ngày nhận: 17/04/2022

Ngày nhận lại: 26/04/2022

Duyệt đăng: 27/04/2022

Từ khóa:

học máy; mô hình định giá tài sản vốn (CAPM); thuật toán hồi quy vector hỗ trợ (SVR); Spark MLlib

TÓM TẮT

Mô hình định giá tài sản vốn (CAPM) lượng hóa mối quan hệ tuyến tính giữa lợi nhuận và rủi ro hệ thống của các tài sản rủi ro. CAPM là một trong những nền tảng lý thuyết của ngành tài chính hiện đại. Tuy nhiên, tính thực nghiệm của CAPM là một chủ đề gây tranh luận đối với các nhà nghiên cứu bởi vì CAPM sử dụng rất nhiều giả định mà khó có thể được đáp ứng trong thực tế. Xu hướng kết hợp trí tuệ nhân tạo và lý thuyết nền tảng tài chính đã tạo ra nhiều mô hình dự báo hiệu quả và phù hợp hơn trong thực nghiệm. Nghiên cứu này thực hiện nhằm 02 mục tiêu chính: Sử dụng thuật toán Support Vector Regression (SVR) trên nền tảng CAPM để dự báo tỷ suất sinh lời của các cổ phiếu riêng lẻ và xác định các yếu tố tác động đến sai số trong dự báo của mô hình kết hợp này. Nghiên cứu sử dụng dữ liệu của các công ty niêm yết trên thị trường chứng khoán Thành phố Hồ Chí Minh giai đoạn từ tháng 12/2012 đến tháng 09/2020, chu kỳ theo tháng. Nghiên cứu chia dữ liệu thành 02 giai đoạn: giai đoạn 01 sử dụng để tối ưu hóa các tham số và giai đoạn còn lại được sử dụng để đánh giá sai số của mô hình dựa trên Spark MLlib. Nghiên cứu chỉ ra rằng mô hình dự báo tỷ suất sinh lời của cổ phiếu sử dụng thuật toán SVR hiệu quả hơn so với CAPM; hơn nữa, nghiên cứu cũng phát hiện ra rằng yếu tố rủi ro đặc thù công ty (VAR), rủi ro tổng thể (SD), sai số của CAPM (RMSECAPM) và tỷ suất sinh lời trung bình (MEAN) là các yếu tố ảnh hưởng đến sự khác biệt giữa sai số dự báo của mô hình SVR đối với từng cổ phiếu đơn lẻ.

ABSTRACT

The Capital Asset Pricing Model (CAPM) measures the linear connection between risky asset return and systematic risk. CAPM is a theoretical underpinning for contemporary finance. The empirical character of the CAPM, on the other hand, is a contentious subject among scholars since the CAPM makes several

assumptions that are difficult to satisfy in reality. In practice, the trend of mixing artificial intelligence with financial foundations theory has resulted in more efficient and appropriate forecasting models. The primary goals of this research are as follows: Using the CAPM and the Support Vector Regression algorithm (SVR), anticipate the return of individual stocks and identify the elements influencing the prediction inaccuracy of this combined model. The analysis makes use of data from firms listed on the Ho Chi Minh City Stock Exchange from December 2012 to September 2020, on a monthly period. The data is divided into two stages in the study: the first is used to optimize the parameters, and the second is used to assess the error of the model based on Spark MLlib. According to research, the stock return forecasting model based on the SVR algorithm is more effective than the CAPM; additionally, the study discovered that company-specific risk (VAR), overall risk (SD), CAPM error (RMSECAPM), and mean return (MEAN) are the main factors influencing the difference between the forecast error of the SVR model for each individual stock.

Keywords:

machine learning; Capital Asset Pricing Model (CAPM); Support Vector Regression; Spark MLlib

1. Giới thiệu

Một trong những nhiệm vụ quan trọng nhưng khó khăn nhất sử dụng chuỗi thời gian là dự báo thị trường chứng khoán (Chen, Xiao, Sun, & Wu, 2017). Dữ liệu chuỗi thời gian về giá chứng khoán thông thường là chuỗi không dừng và rất khó xác định (Tay & Cao, 2001; Zhang, Lin, & Shang, 2017) bởi vì chúng là những chuỗi ngẫu nhiên có xu hướng phi tuyến tính do bị ảnh hưởng bởi nền kinh tế chung, đặc điểm của các ngành, chính trị và thậm chí là tâm lý của các nhà đầu tư (Chen & ctg., 2017; Zhong & Enke, 2017). Giả thuyết thị trường hiệu quả (Efficient Market Hypothesis) cho rằng giá của chứng khoán là một bước đi ngẫu nhiên (Random Walk), do đó khó có thể đoán trước được (Fama, 1970, 1991); mặc dù việc nghiên cứu các mô hình dự báo tỷ suất sinh lợi vẫn đang thu hút rất nhiều sự quan tâm từ giới học thuật và thực nghiệm (Weng, Ahmed, & Megahed, 2017). Nghiên cứu của Atsalakis và Valavanis (2009); Kumar và Thenmozhi (2014); Malkiel (2003) đã nêu ra bằng chứng trái ngược nhau về tính hiệu quả của thị trường tài chính. Các nghiên cứu gần đây đã đề xuất các mô hình nhằm tăng hiệu quả dự báo dựa trên dữ liệu lịch sử. Những phương pháp phổ biến được sử dụng để dự báo kết quả như chỉ báo trung bình động, mô hình tự hồi quy, phân tích khác biệt và mối tương quan (Kumar & Thenmozhi, 2014; Wang, Wang, Zhang, & Guo, 2012). Gần đây hơn, một xu thế mới được tập trung nghiên cứu trong việc dự đoán chuỗi thời gian là học máy, nhằm xử lý dữ liệu ngẫu nhiên và phi tuyến tính (Chen & ctg., 2017).

Nền tảng mô hình định giá tài sản vốn (CAPM) được đề xuất từ những năm 1960 dựa trên lý thuyết về đa dạng hóa và lý thuyết quản lý danh mục đầu tư của Markowizt (Bui & Thai, 2021; Treynor, 1961). Mô hình CAPM lượng hóa mối quan hệ tuyến tính giữa rủi ro hệ thống và lợi nhuận kỳ vọng của các tài sản rủi ro. Mô hình CAPM theo phiên bản của Sharpe-Lintner-Black đã là một công cụ quản lý tài sản quan trọng trong những năm gần đây nhờ lợi thế là đơn giản và dễ sử dụng. Mặc dù vậy, việc sử dụng CAPM trong thực tiễn còn gây nhiều tranh cãi. Những nghiên cứu đầu tiên về CAPM đã minh chứng tồn tại mối quan hệ tuyến tính giữa tỷ số sinh lời và rủi ro hệ thống beta (Black, 1972; Bui & Tran, 2021). Một số nghiên cứu phát hiện đường thị trường chứng khoán khá phẳng, cũng là một thách thức đối với khung lý thuyết CAPM (Amihud, Christensen, & Mendelson, 1992; Breen & Korajczyk, 1993; Fama & French, 2021;

Jagannathan & McGrattan, 1995). Bên cạnh các nghiên cứu ủng hộ lý thuyết CAPM, lại có nhiều nghiên cứu phủ nhận tính thực tiễn của mô hình này (Banz, 1981; Basu, 1983; Chaudhary, 2017; Fama & James, 1973; Lohano & Kashif, 2018). Mặc dù có nhiều ý kiến trái chiều, CAPM cũng đã trở thành một khung lý thuyết nền tảng lý thuyết trong lĩnh vực tài chính hiện đại, hơn nữa, nó cũng được sử dụng phổ biến trong thực nghiệm. Trong một cuộc khảo sát với sự tham gia của hơn 400 CFOs, 75% trong số đó thừa nhận họ sử dụng CAPM để xác định tỷ suất sinh lời kỳ vọng của thị trường đối với các khoản đầu tư cổ phiếu (Graham & Harvey, 2001).

Học máy (Machine Learning) là một phần của ngành khoa học dữ liệu. Thuật ngữ "học máy" đề cập đến lĩnh vực nghiên cứu tập trung vào việc sử dụng các mô hình để đưa ra dự báo. Để xử lý khối lượng lớn dữ liệu, có sẵn các công cụ cho phép phân phối các tác vụ tính toán giữa các nút khác nhau trong một cụm máy tính, để khối lượng công việc được cân bằng và thời gian xử lý giảm xuống. Về vấn đề này, các công cụ như Apache Hadoop hoặc Apache Spark cho phép các thuật toán được chạy theo mô hình phân tán, giúp nhà phát triển tránh được tất cả những bất tiện mà điều này gây ra, chẳng hạn như đồng bộ hóa, truyền dữ liệu và khả năng chịu lỗi, ... Đặc biệt, Apache Spark có thư viện Spark ML, chứa việc triển khai một số thuật toán học máy như mạng nơ-ron, cây quyết định, Random Forest, hồi quy, máy véc-tơ hỗ trợ (SVM) và các thuật toán khác. Kỹ thuật hồi quy vector hỗ trợ (SVR) đã dự báo được lượng mây và sản lượng điện trong hệ thống năng lượng mặt trời tại Nhật Bản. Kết quả dự báo rất khả quan, sai số trung bình bình phương (Root Mean Squared Erorr - RMSE) chỉ khoảng 10% và sai số tuyệt đối (Mean Absolute Error - MAE) xấp xỉ 6% (da Silva Fonseca & ctg., 2012). Một nghiên cứu liên quan đến ngành năng lượng đã sử dụng thuật toán SVR với hàm kernel mũ để dự báo lượng điện của máy phát và so sánh với giá trị thực tế, kết quả dự báo rất tốt (Ramedani, Omid, Keyhani, Shamshirband, & Khoshnevisan, 2014). Phân tích thực nghiệm chỉ ra rằng mô hình SVR với hàm kernel mũ có khả năng dự báo tốt hơn. Cách tiếp cận SVR trong máy học để ước tính chi tiêu mua máy bay quân sự, sử dụng hàm kernel dạng mũ, đã cho kết quả đáng kinh ngạc: sai số trung bình tối thiểu (MSE) là 5.37%, và R^2 là 99%, một kết quả tốt ngoài kỳ vọng (Tong, 2015). Ứng dụng máy học trong lĩnh vực tài chính khá đa dạng, ví dụ như nghiên cứu việc sử dụng mô hình Fama 03 và 05 nhân tố (Gogas, Papadimitriou, & Karagkiozis, 2018). Các tác giả đã so sánh SVR với phương pháp OLS trong mô hình CAPM, mô hình Fama 03 và 05 nhân tố, cũng như trong mô hình lý thuyết kinh doanh chênh lệch giá (APT), sử dụng dữ liệu từ thị trường chứng khoán Mỹ cho mô hình Fama 03 nhân tố với 1062 quan sát (07/1926 - 12/2014), mô hình Fama 05 nhân tố với 618 quan sát (07/1963 - 12/2014), và mô hình APT với 346 quan sát (02/1986 - 12/2014). Hệ số R^2 hiệu chỉnh và MAPE được sử dụng để đo lường chất lượng dự báo của mô hình. Theo kết quả nghiên cứu, phương pháp sử dụng SVR với hàm kernel dạng mũ và dạng đa thức đã tỏ ra vượt trội so với phương pháp hồi quy OLS truyền thống khi xét tới MAPE và hệ số R^2 hiệu chỉnh. Henrique, Sobreiro, và Kimura (2018) đã sử dụng SVR để ước tính giá cổ phiếu theo ngày với các mô hình được hiệu chỉnh theo thời gian.

Bộ dữ liệu NASDAQ-100 được Abraham, Nath, và Mahanti (2001) sử dụng, các tác giả này đã tiên phong sử dụng máy học trong các nghiên cứu thực nghiệm về thị trường chứng khoán. Các thuật toán được đưa ra so sánh bao gồm Phân tích thành phần chính (PCA), Mạng thần kinh nhân tạo (ANN), và Mạng thần kinh bóng mờ tiến hóa (NFUZZ) (Abraham & ctg., 2001). Tương tự, ANN và thuật toán Máy vector hỗ trợ (SVM) được sử dụng tại thị trường Chicago dành cho 05 hợp đồng tương lai, và tác giả sử dụng Sai số bình phương trung bình chuẩn hóa (NMSE), Độ cân xứng có hướng (DS), và Sai số tuyệt đối trung (MAE) (Cao & Tay, 2003). Gần đây, kỹ thuật Hồi quy vector hỗ trợ (SVR) được sử dụng để dự báo giá vàng (Yuan, Lee, & Chiu, 2020), và thuật toán Di truyền - Hồi quy vector hỗ trợ bình phương nhỏ nhất (GA-LSSVR) được sử dụng để kiểm định độ nhạy và đánh giá chất lượng mô hình thông qua chỉ số

MAPE. Kỹ thuật SVR có thể phát hiện mối quan hệ phi tuyến tính mà phương pháp OLS không thực hiện được. Tại Việt Nam, K. T. Tran, Banh, và Nguyen (2012) đã kết hợp giải thuật di truyền và SVR để dự đoán giá cổ phiếu trên thị trường chứng khoán Việt Nam; Trinh (2013) đã ứng dụng kỹ thuật học máy SVR để xây dựng được chương trình dự đoán xu hướng tăng giảm của cổ phiếu dựa theo dữ liệu từ tập dữ liệu Twitter. Do đó, nghiên cứu sử dụng SVR dựa trên mô hình CAPM để dự báo tỷ suất sinh lời của cổ phiếu đơn lẻ, đồng thời xác định các yếu tố ảnh hưởng đến sự khác biệt sai số dự báo tỷ suất sinh lời đối với từng cổ phiếu đơn lẻ tại Việt Nam đang rất hạn chế. Do đó, thông qua sử dụng Spark MLlib, nghiên cứu này tận dụng các lợi thế của mô hình CAPM cùng với tính hiệu quả của thuật toán SVR bằng việc kết hợp CAPM và SVR, qua đó tạo ra kết quả dự báo chính xác hơn so với các nghiên cứu trước đó nhờ vào tính ưu việt cũng như mức độ phổ biến của SVR. Mô hình kết hợp này được xem như là một phương pháp thay thế mô hình CAPM truyền thống. Lợi thế của mô hình này là khả năng "học" để cải thiện độ chính xác thông qua việc sử dụng thuật toán máy học, kiểm soát nhiễu, khám phá các thành phần ẩn của dữ liệu, và ước tính các hàm phi tuyến. Mô hình có sử dụng SVR đã tỏ ra vượt trội cách tiếp cận CAPM truyền thống nhờ vào các điểm này.

Ngoài phần giới thiệu, thì bố cục của bài báo như sau, tổng quan lý thuyết về mô hình CAPM và thuật toán SVR được trình bày trong phần 2, phương pháp nghiên cứu được giải thích trong phần 3, và kết quả thực nghiệm được đưa ra trong phần 4. Cuối cùng, phần kết luận của bài báo này được chỉ ra trong phần 5 của bài báo.

2. Cơ sở lý thuyết

2.1. Mô hình CAPM

CAPM là một tập hợp các ước tính tỷ suất sinh lời kỳ vọng của các tài sản rủi ro ở trạng thái cân bằng. Nó được hình thành trên nền tảng lý thuyết lựa chọn danh mục đầu tư (Markowitz, 1952; H. T. Tran, 2020). Các giả định của mô hình bao gồm:

- Các nhà đầu tư là e ngại rủi ro và luôn chọn danh mục trung bình - phương sai - hiệu quả.
- Thời gian nắm giữ danh mục chỉ trong một kỳ đơn lẻ.
- Kỳ vọng của các nhà đầu tư là thuần nhất.
- Tất cả các tài sản đều công khai, giao dịch đại chúng, có thể chia nhỏ tùy ý và cho phép bán khống.
- Các nhà đầu tư có thể vay và cho vay một lượng tùy ý ở mức lãi suất phi rủi ro.
- Thông tin là có sẵn và công khai.
- Không có thuế và chi phí giao dịch.

Nghiên cứu này bắt đầu với một nhà đầu tư đặt ra tỷ trọng α đối với tài sản thứ i và $1-\alpha$ đối với danh mục thị trường ($0 \le \alpha \le 1$). Khi đó tỷ suất sinh lợi là một hàm theo α như sau:

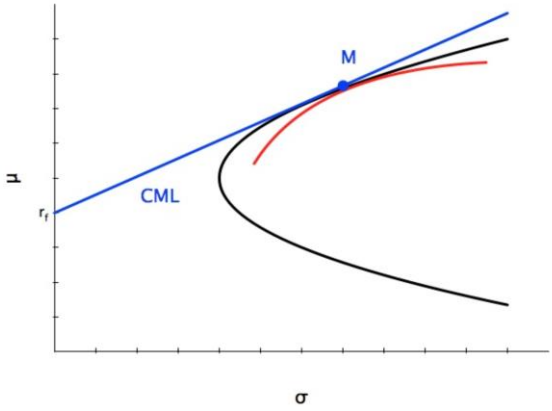
$$r(\alpha) = \alpha r_i + (1-\alpha) r_M \quad (1)$$

Tính giá trị kỳ vọng và phương sai ta thu được kết quả như sau:

$$E\big(r(\alpha)\big) = \alpha E(r_i) + (1-\alpha) E(r_M) \quad (2)$$

$$\sigma^2\big(r(\alpha)\big) = \alpha^2 \sigma^2(r_i) + (1-\alpha)^2 \sigma^2(r_M) + 2\alpha(1-\alpha)\, cov(r_i, r_M) \quad (3)$$

Khi α thay đổi, làm cho các điểm $\left(\sigma\left(r(\alpha)\right), E\left(r(\alpha)\right)\right)$thay đổi trên đường cong màu đỏ và chỉ cắt đường biên hiệu quả tại điểm M ứng với α = 0 (Hình 1).



Hình 1. Đường thị trường vốn (Capital Market Line, CML)
và các cơ hội đầu tư khi α thay đổi

Khi tất cả các nhà đầu tư đều có cùng kỳ vọng, họ cùng chọn một danh mục sao cho tối ưu hóa tỉ số Sharpe: $\underset{\alpha}{Max}\frac{E(r(\alpha))-r_f}{\sigma(r(\alpha))}$. Kết quả là danh mục M được lựa chọn (ứng với α = 0).

Vì đường CML tiếp xúc với đường màu đỏ tại M nên hệ số góc của cả hai đường phải bằng nhau. Để đơn giản hơn ta đặt $g(\alpha) = E\left(r(\alpha)\right), h(\alpha) = \sqrt{\sigma^2\left(r(\sigma)\right)}$ và f là hàm biểu thị mối quan hệ của $\sigma\left(r(\alpha)\right)$và $E\left(r(\alpha)\right)$ trên đường cong màu đỏ, tức là $E\left(r(\alpha)\right) = f\left(\sigma\left(r(\alpha)\right)\right)$. Khi đó: $g(\alpha) = f\left(h(\alpha)\right)$. Lấy đạo hàm ta được: $g'(\alpha) = f'\left(h(\alpha)\right)h'(\alpha)$. Do đó: $f'\left(h(\alpha)\right) = \frac{g\prime(\alpha)}{h\prime(\alpha)}$. Tính toán đạo hàm g và h như sau:

$$g(\alpha) = \alpha E(r_i) + (1-\alpha)E(r_M) \Rightarrow g'(\alpha) = E(r_i) - E(r_M) \quad (4)$$

$$h^2(\alpha) = \alpha^2\sigma^2(r_i) + (1-\alpha)^2\sigma^2(r_M) + 2\alpha(1-\alpha)\,cov(r_i, r_M)$$
$$\Rightarrow 2h(\alpha)h'(\alpha) = 2\alpha\sigma^2(r_i) + 2(\alpha-1)\sigma^2(r_M) + 2(1-2\alpha)\,cov(r_i, r_M)$$
$$\Rightarrow h'(\alpha) = \frac{2\alpha\sigma^2(r_i) + 2(\alpha-1)\sigma^2(r_M) + 2(1-2\alpha)\,cov(r_i, r_M)}{2h(\alpha)} \quad (5)$$

Với α = 0 ta có hệ số góc của đường con màu đỏ là:

$$h'(\alpha) = \frac{E(r_i) - E(r_M)}{\frac{-2\sigma^2(r_M) + 2\,cov(r_i, r_M)}{2\sqrt{\sigma^2(r_M)}}} = \frac{\left(E(r_i) - E(r_M)\right)\sigma(r_M)}{cov(r_i, r_M) - \sigma^2(r_M)} \quad (6)$$

Mặt khác, đường thẳng CML đi qua điểm $\left(0, r_f\right)$và điểm $M\left(\sigma(r_M), E(r_M)\right)$ nên có hệ số góc là: $\frac{E(r_M)-r_f}{\sigma(r_M)}$. Vì tại M đường cong màu đỏ tiếp xúc với đường CML nên hệ số góc cả hai đường bằng nhau. Do đó: $\frac{\left(E(r_i)-E(r_M)\right)\sigma(r_M)}{cov(r_i,r_M)-\sigma^2(r_M)} = \frac{E(r_M)-r_f}{\sigma(r_M)}$ biến đổi tương đương ta thu được: $E(r_i) - r_f = \frac{cov(r_i,r_M)}{\sigma^2(r_M)}\left(E(r_M) - r_f\right)$. Đặt $\beta_i = \frac{cov(r_i,r_M)}{\sigma^2(r_M)}$ ta thu được công thức mô hình CAPM quen thuộc như sau: $E(r_i) = r_f + \beta_i\left(E(r_M) - r_f\right)$ (7)

Rủi ro hệ thống của một chứng khoán thị trường được đo bởi hệ số beta, hệ số này đo

lường mức độ đóng góp của một cổ phiếu vào biến động tỷ suất sinh lời của cả danh mục. Hệ số beta của các chứng khoán được tính với độ dài khoảng thời gian 24 tháng.

2.2. Thuật toán hồi quy vector hỗ trợ (SVR)

Phương pháp phân lớp dựa vào thuật toán Support Vector Machine (SVM) là ánh xạ từ các biến độc lập với N quan sát tới một không gian một hoặc nhiều chiều nhằm phân lớp giữa các nhóm. Phương pháp này được đề xuất bởi Vapnik, sử dụng tập huấn luyện $\{(x_i, y_i)\}_{i=\overline{1,N}}$ để xây dựng mô hình tuyến tính với biên phân lớp phi tuyến. Phân lớp giữa các nhóm được thực hiện bằng cách sử dụng siêu phẳng tối ưu được tính toán dựa vào N quan sát, trong đó **x** là biến độc lập, **y** là biến phân loại ($y_i \in \{-1,1\}$). Do đó, siêu phẳng phân lớp được cho bởi phương trình : $H: w^T\Phi(x_k) + b = 0$ (8), trong đó $\Phi: R^n \to R^m$ là một ánh xạ từ tập dữ liệu gốc tới không gian chiều cao hơn để hỗ trợ việc phân loại.

Nghiên cứu này giả định rằng khoảng cách ngắn nhất giữa các điểm tới siêu phẳng (H) bằng 1 đối với cả hai lớp nhờ điều chỉnh trọng số w và hệ số b. Bài toán SVM là ước lượng các tham số *w, b* theo phương pháp này.

Giả sử siêu phẳng H có thể phân loại tập dữ liệu một cách hoàn hảo; từ đó, $y_k[w^T\Phi(x_k) + b] \geq 1$ (9) ∀k = 1,2,…, N, và các tham số tối ưu trong mô hình được ước tính bằng cách tìm cực tiểu của hàm mục tiêu $\|w\|$ theo giá trị của w và b, điều kiện phân loại tương ứng là dấu của hàm $h(x) = w^T\Phi(x) + b$ (Vapnik, 2013). Mặc dù vậy, khó để tìm được ánh xạ Φ có thể chia tách hoàn hảo. Cortes và Vapnik (1995) đề xuất một ý tưởng mới cho phép gán sai tên một số quan sát, phương pháp này sử dụng biến bù ξ_i để đo lường sai số của quan sát thứ *i*. Bài toán SVM trở thành: $\min_{w,b,\xi}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i\right)$ (10),

với $y_i[w^T\Phi(x_i) + b] \geq 1 - \xi_i, \xi_i \geq 0$ (11)

Thuật toán SVR dựa trên ý tưởng tương tự như SVM, ngoại trừ việc biến phụ thuộc là một biến liên tục theo giá trị thực. Tuy nhiên, SVR sử dụng hàm hồi quy là một siêu phẳng như (11) (Patel, Shah, Thakkar, & Kotecha, 2015; Qu & Zhang, 2016). Đường biên được mô tả dưới dạng:

$$|y - f(x,w)|_\varepsilon = \begin{cases} 0, |y - f(x,w)| \leq \varepsilon \\ |y - f(x,w)| - \varepsilon, |y - f(x,w)| > \varepsilon \end{cases} \quad (12)$$

Phương pháp SVR tìm cực tiểu của R theo ε và $\|w\|^2$ trong phương trình: $R = \frac{1}{2}\|w\|^2 + C(\sum_{i=1}^{N}|y - f(x_i,w)|_\varepsilon)$ (13) với C là siêu tham số.

3. Phương pháp nghiên cứu

Spark là một công cụ hàng đầu trong Hệ sinh thái Hadoop. MapReduce với Hadoop chỉ có thể được sử dụng để xử lý hàng loạt và không thể hoạt động trên dữ liệu thời gian thực. Spark có thể hoạt động độc lập hoặc trên khuôn khổ Hadoop để tận dụng dữ liệu lớn và thực hiện phân tích dữ liệu thời gian thực trong môi trường máy tính phân tán. Học máy là một trong những ứng dụng chính của Spark. Spark MLlib bao gồm các thuật toán học máy phổ biến để hồi quy, phân loại, phân cụm, lọc cộng tác và khai thác mẫu thường xuyên. Nó cũng cung cấp một loạt các tính năng để xây dựng đường ống (pipelines), lựa chọn và điều chỉnh mô hình, cũng như lựa chọn, khai thác và chuyển đổi. Các phiên bản đầu tiên của Spark MLlib chỉ bao gồm một giao diện lập trình ứng dụng (Application Programming Interface - API) dựa trên bộ dữ liệu phân tán linh hoạt (Resilient Distributed Dataset - RDD). API dựa trên DataFrame hiện là API chính cho Spark. API dựa trên DataFrames giúp dễ dàng chuyển đổi các tính năng bằng cách cung cấp tính trừu

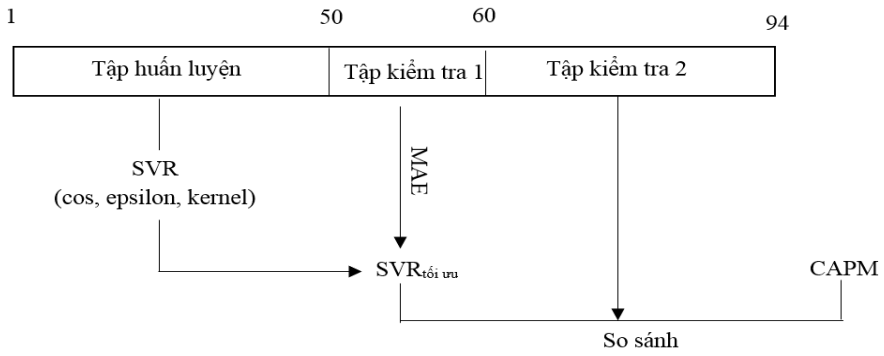
tượng cấp cao hơn để biểu diễn dữ liệu dạng bảng tương tự như bảng cơ sở dữ liệu quan hệ, làm cho nó trở thành một lựa chọn tự nhiên để triển khai các đường ống. Nghiên cứu này sử dụng phiên bản Hadoop và Spark ML phiên bản 3.1.1.

Nghiên cứu xây dựng một hệ thống phân tích cổ phiếu để dự đoán mức tăng hàng ngày trên thị trường chứng khoán dựa trên dữ liệu cafef.vn, và vn.investing.com hoặc các tài nguyên trực tuyến khác. Nghiên cứu này thu thập dữ liệu từ Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh (Ho Chi Minh Stock Exchange, HOSE), nghiên cứu đã loại những cổ phiếu niêm yết sau 12/2012 và hủy niêm yết trước tháng 09/2020. Do đó, dữ liệu nghiên cứu gồm giá cổ phiếu đóng cửa điều chỉnh của 212 cổ phiếu, và lãi suất trái phiếu chính phủ kỳ hạn 01 năm từ tháng 12/2012 đến tháng 09/2020 (gồm 94 tháng). Nghiên cứu chia dữ liệu thu được hàng ngày thành tập dữ liệu đào tạo và thử nghiệm để dự đoán những cổ phiếu có mức tăng hàng ngày cao bằng cách sử dụng mô-đun học máy của Spark và sau đó dự đoán mối tương quan giữa giá cổ phiếu dựa trên các hệ số trong mô hình hồi quy. Theo đó, dữ liệu được xử lý bằng cách xóa bỏ các dữ liệu bị thiếu và các dữ liệu ngoại lai. Bảng 1 mô tả các biến.

Nghiên cứu sử dụng mô hình CAPM kết hợp thuật toán SVR. Qua đó, quy trình nghiên cứu của dự án được xây dựng như Hình 2 và gồm 02 bước:

Bước 1: Đối với nhóm huấn luyện, nghiên cứu sử dụng 50 tháng đầu tiên (có nghĩa 50 * 212 dữ liệu) để làm tập huấn luyện. Có tổng cộng 60 mô hình kiểm định với các quan sát này (cost = 1, 0.5, 0.1, 0.05, 0.01, 0.001; epsilon = 1, 0.8, 0.6, 0.2, 0.1 và kernel = linear, radial, polynomial). Các quan sát từ 51 đến 60 được sử dụng như tập kiểm tra 01 nhằm lựa chọn mô hình có sai số MAE thấp nhất trong 60 mô hình trên.

Bước 2: Sử dụng các quan sát từ 61 - 94 để làm tập kiểm tra 02 nhằm đánh giá hiệu quả của việc kết giữa SVR và CAMP so mô hình CAPM gốc để chỉ ra tính hiệu quả của thuật toán. Công thức xác định kết quả đầu ra của mô hình SVR là: $r_{it} = r_f + f(premium_{it})$, với f là hàm xác định bởi thuật toán SVR nhờ tham số xác định trong Bước 1. Sau khi xác định kết quả dự báo của cả 02 mô hình, nghiên cứu đã tính toán độ lệch giữa kết quả ước lượng với giá trị trên thực tế. Cuối cùng, kiểm định Wilcoxon được sử dụng để xác định tính hiệu quả của mô hình SVR so với mô hình CAPM.



Hình 2. Quy trình nghiên cứu

Để đánh giá các nhân tố tác động đến sai số của mô hình SVR, nghiên cứu này sử dụng hồi quy dữ liệu chéo theo phương trình (1) theo nguyên cứu trước đó của H. T. Tran (2020). Phương trình (14) hàm ý rằng sai số trong mô hình SVR sẽ phụ thuộc vào sai số của mô hình lý thuyết nền CAPM (RMSECAPM) và các nhân tố có liên qua như đặc trưng rủi ro tổng thể (SD), rủi ro đặc thù công ty (VAR), tỷ suất sinh lợi kỳ vọng (MEAN) và đặc trưng rủi ro hệ thống (BETA). Phương trình hồi quy có dạng:

$$\text{RMSESVR}_i = \beta_0 + \beta_1\text{RMSECAPM}_i + \beta_2\text{VAR}_i + \beta_3\text{SD}_i + \beta_4\text{MEAN}_i + \beta_5\text{BETA}_i + \varepsilon_i \quad (14)$$

Bảng 1

Mô tả các biến

Biến	Công thức	Mô tả
r_{Mt}	$r_{Mt} = \frac{Vnindex_t - Vnindex_{t-1}}{Vnindex_{t-1}} \times \frac{365}{n}$	Tỷ suất sinh lời của danh mục đầu tư thị trường (%/năm)
r_{it}	$r_{it} = \frac{price_{it} - price_{i\,(t-1)}}{price_{i\,(t-1)}} \times \frac{365}{n}$	Tỷ suất sinh lời của cổ phiếu thứ *i* tại thời điểm t (%/năm)
r_{ft}		Lợi suất của trái phiếu chính phủ kỳ hạn 01 năm ở thời điểm t
β_{it}	$\beta_{it} = \frac{cov(r_{it}, r_{Mt})}{var(r_{Mt})}$	Hệ số beta của cổ phiếu thứ *i* tại thời điểm t (beta được ước tính với dữ liệu 24 tháng)
premium_{it}	$premium_{it} = \beta_{it}(r_{Mt} - r_{ft})$	Phần bù rủi ro
RCAPM_{it}	$\text{RCAPM}_{it} = r_{ft} + \text{premium}_{it}$	Tỷ suất sinh lời kỳ vọng của cổ phiếu thứ *i* tại thời điểm t theo ước tính của CAPM
RSVR_{it}	$\text{RSVR}_{it} = r_{ft} + f(\text{premium}_{it})$	Tỷ suất sinh lời kỳ vọng của cổ phiếu thứ *i* tại thời điểm t theo ước tính của SVR
DCAPM_{it}	$\text{DCAPM}_{it} = \lvert r_{it} - \text{RCAPM}_{it} \rvert$	Sai số tuyệt đối của dự báo theo CAPM
DSVRit	$\text{DSVR}_{it} = \lvert r_{it} - \text{RSVR}_{it} \rvert$	Sai số tuyệt đối của dự báo theo SVR
RMSE	$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(R_t - F_t)^2}$	Đo lường độ lệch của giá trị dự báo so với giá trị thực tế
MAE	$MAE = \frac{1}{n}\sum_{t=1}^{n}\lvert R_t - F_t \rvert$	
RMSESVR_i	$RMSESVR_i = \frac{1}{T}\sum_{t=1}^{T} DSVR_{it}$	RMSE của cổ phiếu thứ *i* trong mô hình SVR

Biến	Công thức	Mô tả
$RMSECAPM_i$	$RMSESVR_i = \frac{1}{T}\sum_{t=1}^{T} DSVR_{it}$	RMSE của cổ phiếu thứ *i* trong mô hình CAPM
VAR_i	$var(\varepsilon_{it})$	Phương sai của phần dư trong hồi quy chuỗi thời gian $r_{it} - r_{ft} = \beta_0 + \beta_i(r_{Mt} - r_{ft}) + \varepsilon_{it}$
$MEAN_i$	$MEAN_i = \frac{1}{T}\sum_{t=1}^{T} r_{it}$	Tỷ suất sinh lời trung bình của cổ phiếu thứ *i*
$BETA_i$	$BETA_i = \frac{cov(r_{it}, r_{Mt})}{var(r_{Mt})}$	Hệ số beta của cổ phiếu thứ *i* tại bước 2.
SD_i	$SD_i = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(r_{it} - MEAN_i)^2}$	Độ lệch chuẩn của cổ phiếu thứ *i*

Nguồn: Tác giả tổng hợp

4. Kết quả nghiên cứu

Trong suốt giai đoạn 12/2012 - 09/2020, hệ số beta và MEAN khá ổn định; hệ số beta biến động từ -0.641 tới 0.867, hơn một nửa quan sát là giá trị âm. Trong khi đó, MEAN dao động từ -0.387 tới 0.415. Hầu hết các cổ phiếu có MEAN dương. Dữ liệu cụ thể được tổng hợp trong Bảng 2.

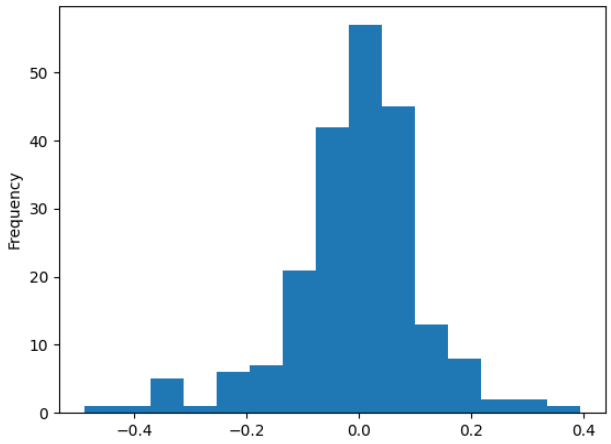
Bảng 2

Thống kê mô tả

Biến	Trung bình	Độ lệch chuẩn	Nhỏ nhất	Trung vị	Lớn nhất
BETA	-0.006	0.117	-0.489	0.002	0.394
VAR	4.680	1.303	0.103	2.143	229.493
MEAN	0.174	0.327	-0.482	0.120	2.810
rf	0.041	0.017	0.003	0.041	0.086
VNindex	0.135	0.703	-2.840	0.153	2.025

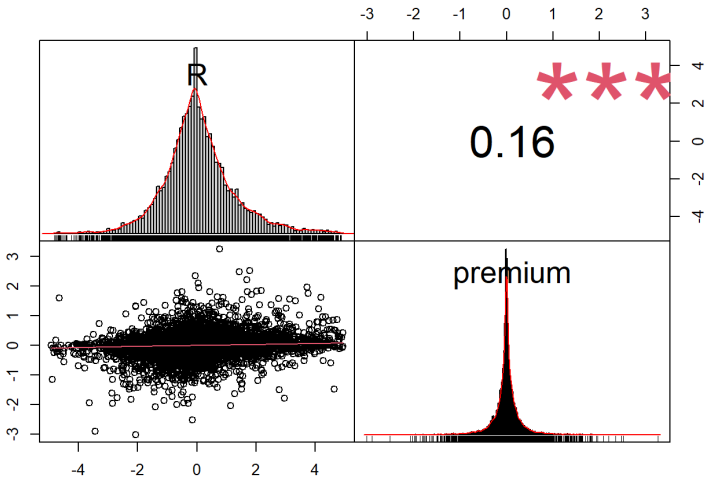
Nguồn: Kết quả phân tích của tác giả

Phân phối của beta và MEAN được thể hiện bằng biểu đồ histogram (Hình 3). Phân phối của beta có dạng đối xứng với trung bình và trung vị bằng xấp xỉ 0; hai giá trị ngoại biên nằm xa tương đối so với các quan sát còn lại. Cổ phiếu KSB có hệ số beta thấp nhất -0.489 và tỷ suất trung bình tương ứng là -0.387 (-38.7 %/năm). Ngoài ra, cổ phiếu KSB cũng là cổ phiếu có MEAN thấp nhất trong số 212 công ty. Ở chiều ngược lại, beta của SVC là 0.394, cao nhất trong 212 mã cổ phiếu. Nhìn chung, biến động về tỷ suất sinh lời của tất cả các cổ phiếu đều thấp hơn thị trường (tất cả beta đều nhỏ hơn 1). Lý do có thể bắt nguồn từ việc chuỗi quan sát quá dài, làm mất đi tính nhạy cảm vốn có, hoặc do tỷ suất sinh lời của VNIndex không phải là một đại diện tốt cho tỷ suất sinh lời của danh mục thị trường.



Hình 3. Phân phối hệ số beta

Hình 4 chỉ ra biểu đồ phân phối của R ($R = r - r_f$) và phần bù rủi ro phân tán đều xung quanh 0. Tương quan giữa R và phần bù rủi ro có ý nghĩa thống kê ở mức 5% nhưng độ lớn khá yếu (chỉ 0.16). Kết quả này hàm ý rằng biến động của tỷ suất sinh lời vượt trội được giải thích một phần bởi phần bù rủi ro theo mô hình CAPM. Biểu đồ phân tán trong Hình 3 cho thấy có 02 kernel phù hợp, là dạng tuyến tính và dạng hàm mũ. Vì vậy, tham số cần kiểm định bao gồm cả dạng kernel hàm tuyến tính và hàm mũ, với cost nhận các giá trị 1, 0.5, 0.1, 0.05, 0.01, 0.001 và epsilon nhận giá trị 1, 0.8, 0.6, 0.2, 0.1. Tổng cộng có 60 mô hình kiểm định được thực hiện, cho kết quả hàm kernel dạng căn có cost là 1, epsilon là 0.1, với MAE thấp nhất là 0.8833, và hàm kernel dạng tuyến tính có cost là 0.001, epsilon là 0.1 với MAE nhỏ nhất là 0.8834. Mặc dù kernel dạng mũ có MAE tốt hơn so với dạng tuyến tính, chênh lệch là không đáng kể; hơn nữa, kernel tuyến tính có tốc độ xử lý nhanh hơn và vượt trội so với dạng hàm mũ.



Hình 4. Phân phối và tương quan giữa R và phần bù rủi ro

Kết quả dự báo theo SVR và CAPM được tổng hợp theo chỉ số MAE và RMSE của 212 cổ phiếu trong Bảng 3 chỉ ra rằng mô hình SVR dự báo tốt hơn so với CAPM theo tiêu chí MAE và RMSE. Cụ thể là, sai số trung bình của của dự báo SVR là 0.9087, nhỏ hơn so với giá trị tương tự của CAPM là 0.9251. Các giá trị thống kê khác của SVR cũng nhỏ hơn so với CAPM, ngoại trừ giá trị cực tiểu của RMSE.

Bảng 3

Thống kê mô tả MAE và RMSE

Statistic	*Nhỏ nhất*	*Q_1*	*Trung vị*	*Trung bình*	*Q_3*	*Lớn nhất*
			MAE			
SVR	0.3539	0.7065	0.9022	0.9087	1.0876	1.6578
CAPM	0.3499	0.7106	0.9060	0.9251	1.1277	1.6567
			RMSE			
SVR	0.4396	0.9917	1.1965	1.2257	1.4613	2.0174
CAPM	0.4341	0.9830	1.1969	1.2345	1.4571	2.0554

Nguồn: Kết quả phân tích của tác giả

Nghiên cứu sử dụng kiểm định Wilcoxon để so sánh tính hiệu quả của việc dự báo giữa SVR và CAPM. Giả thuyết chính của kiểm định là không có sự khác nhau về sai số dự báo giữa hai mô hình (H_0), và giả thuyết đối là mô hình SVR có sai số nhỏ hơn so với CAPM (H_1). Kết quả kiểm định Wilcoxon. Với p-value = 0.04848, nhỏ hơn mức ý nghĩa 0.05; giả thuyết H_0 bị bác bỏ. Mô hình SVR phù hợp hơn mô hình CAPM. Kết quả này phù hợp với nghiên cứu của Gogas và cộng sự (2018). Tuy nhiên, Gogas và cộng sự (2018) sử dụng danh mục cổ phiếu thay vì cổ phiếu đơn lẻ, nên R^2 có kết quả tốt hơn, trong khoảng từ 0.59 đến 0.75.

Mô hình SVR được xây dựng dựa trên nền tảng của mô hình CAPM, nên khả năng dự báo của SVR sẽ phụ thuộc vào độ chính xác của CAPM. Vì vậy, RMSECAPM sẽ tác động trực tiếp tới sai số dự báo của SVR. Bên cạnh đó, các yếu tố ảnh hưởng tới khả năng dự báo của CAPM được trình bày tại Bảng 3. Hầu hết các biến đều có giá ý nghĩa về mặt thống kê ở mức ý nghĩa 1%, ngoại trừ BETA. Kết quả này hàm ý rằng biến động về giá trị của beta không tác động đến mức độ chính xác trong dự báo của mô hình. Biến SD có ước lượng hệ số hồi quy cao nhất và vượt trội so với các biến còn lại, ngụ ý rằng sai số mô hình SVR phụ thuộc phần lớn vào rủi ro tổng thể của các cổ phiếu. Hệ số ước lượng của SD có giá trị là 0.886 hàm ý rằng nếu các nhân tố tác động khác không đổi, mỗi đơn vị rủi ro tổng thể tăng thêm dự báo rằng sai số RMSESVR tăng thêm 0.886 đơn vị. Hệ số ước lượng của RMSECAPM có giá trị lớn thứ hai (0.1166) và có p-value = 0.000 chứng tỏ sai số của CAPM tác động có ý nghĩa thống kê đến sai số của mô hình SVR. Cụ thể, mỗi đơn vị tăng thêm trong RMSECAPM và giữ nguyên các yếu tố còn lại, ta kỳ vọng RMSESVR tăng thêm 0.1166 đơn vị. Các biến VAR, MEAN đều có hệ số ước lượng dương và có ý nghĩa thống kê (mức 5%) cho thấy chúng đều tác động cùng chiều với biến phụ thuộc. Bên cạnh đó, hệ số xác định R^2 = 0.999, rất cao, cho thấy các biến độc lập giải thích được phần lớn biến động của RMSESVR.

Bảng 4

Kết quả hồi quy RMSESVR

$$\text{RMSESVR}_i = \beta_0 + \beta_1 \text{RMSECAPM}_i + \beta_2 \text{VAR}_i + \beta_3 \text{SD}_i + \beta_4 \text{MEAN}_i + \beta_5 \text{BETA}_i + \varepsilon_i$$

	Beta	**t-value**	**p-value**	**R^2**	**Adj R^2**
β_0	0.0139	2.308	0.022	0.99	0.99
β_1	0.1166	5.551	0.000		
β_2	0.0016	2.997	0.003		
β_3	0.886	39.372	0.000		
β_4	0.0919	10.002	0.000		
β_5	0.0129	0.662	0.509		

Nguồn: Kết quả phân tích của tác giả

5. Thảo luận và kết luận

5.1. Thảo luận

Trong nghiên cứu này, thuật toán SVR đã được sử dụng với các tham số khác nhau để tìm ra tham số phù hợp nhất, cụ thể hàm kernel dạng tuyến tính có cost là 0.001, epsilon là 0.1 với MAE nhỏ nhất là 0.8834. Nghiên cứu này đã kết hợp thuật toán SVR và mô hình CAPM thay vì chỉ sử dụng riêng lẻ CAPM để dự báo tỷ suất sinh lợi của cổ phiếu riêng lẻ. Có tất cả 60 mô hình dự báo từ tập huấn luyện, nên sử dụng để xử lý sẽ đạt hiệu quả cao về mặt thời gian tính toán kết quả.

Thuật toán hồi quy KNN (K-Nearest Neighbors Algorithm) và thuật toán hồi quy vectơ hỗ trợ chuyên sâu epsilon tuyến tính đã được sử dụng để dự đoán giá đóng cửa hàng ngày của các cổ phiếu được chọn của DSE. Việc xác nhận chéo, cùng với quá trình lặp lại, đã được thực hiện để xác định các siêu tham số tối ưu. Chúng tôi đã đưa ra dự đoán sau khi chọn mô hình tốt nhất và áp dụng nó vào dữ liệu thử nghiệm. Nghiên cứu của chúng tôi cho thấy SVR tuyến tính có sai số nhỏ hơn KNN và SVR tuyến tính có giá trị cao hơn và được điều chỉnh R^2 giá trị trong cả bộ kiểm tra và bộ huấn luyện. Vì vậy, SVR tuyến tính hoạt động tốt hơn và nó có thể được sử dụng để dự đoán 01 ngày trước giá đóng cửa của thị trường chứng khoán, cung cấp dữ liệu lịch sử trước đó. Tóm lại, vì thị trường chứng khoán là một lĩnh vực tài chính quan trọng, việc so sánh giữa các mô hình chuỗi thời gian có thể giúp xác định xem nên mua hay bán cổ phiếu và mục đích quan trọng này có thể được phục vụ với sự trợ giúp của nghiên cứu so sánh này về thị trường chứng khoán phỏng đoán. Trong các nghiên cứu trước đây, chúng tôi thấy rằng SVR tốt hơn nhưng không tìm thấy nó là tuyến tính hay phi tuyến và chúng tôi cũng không so sánh với KNN. Chúng tôi nhận thấy rằng SVR tuyến tính tốt hơn KNN. Nghiên cứu này chỉ được thực hiện trên ba công ty được chọn và đã nghiên cứu thêm về hiệu suất của SVR, và KNN nên được kiểm tra cho các dữ liệu chuỗi thời gian khác.

Lựa chọn kernel phụ thuộc vào bản chất của tập dữ liệu, nếu dữ liệu tập trung thành cụm tròn thì kernel radial là phù hợp nhất, nếu dữ liệu phân tán xung quanh một siêu phẳng thì kernel linear phù hợp nhất, kernel polynomial phù hợp cho dữ liệu phân tán theo hàm đa thức. Nghiên cứu này đã sử dụng kernel dạng tuyến tính, điều này phù hợp với mối quan hệ giữa các biến trong mô hình lý thuyết CAPM. Tham số cost đặc trưng cho chi phí sai lệch, cost cao hàm ý cho phép sai lệch lớn (thường dẫn đến hiện tượng underfitting) nhưng cost thấp sẽ ít cho phép sai lệch hơn trong tập huấn luyện và có thể gây ra hiện tượng overfitting. Epsilon là tham số điều chỉnh khoảng cách giữa giá trị thực và giá trị dự báo (khoảng cách này bằng 0 nếu nhỏ hơn epsilon). Tham số gamma cho phép thay đổi hình dạng của hàm mật độ Gausse trong kernel radial.

Sai số dự báo đo lường bằng chỉ số RMSE ở các mã chứng khoán phụ thuộc vào các nhân tố: sai số của mô hình CAPM, rủi ro đặc thù, tỷ suất sinh lợi trung bình và rủi ro tổng thể. Nhân tố sai số của CAPM đo lường bằng RMSECAPM có hệ số vượt trội và cùng chiều với RMSESVR hàm ý rằng sai lệch khi fitted bằng CAPM càng cao kỳ vọng rằng sai lệch khi fitted bằng SVR cũng cao tương ứng. Nhân tố đặc trưng cho rủi ro hệ thống (BETA) một lần nữa không có tác động có ý nghĩa thống kê đến RMSESVR. Khi sử dụng các nhân tố trên giải thích cho sự biến động trong RMSESVR, hệ số R^2 thu được là 0.99 rất cao hàm ý rằng các nhân tố này hầu như đã giải thích hoàn toàn các biến động trong RMSESVR.

5.2. Kết luận

Kết quả phân tích hồi quy đã chứng tỏ các hệ số ước lượng đều dương, nói cách khác, các biến giải thích có tác động cùng chiều với biến phụ thuộc. Kết quả này hàm ý rằng việc kiểm soát các biến độc lập theo hướng giảm kỳ vọng sẽ giảm sai số trong mô hình SVR. Mặc dù CAPM là một lý thuyết nền tảng quan trọng nhưng khả năng ứng dụng trong thực nghiệm còn nhiều tranh

cải do nó có quá nhiều giả định khó đảm bảo trong thực tế. Mô hình kết hợp SVR được đề xuất như một mô hình thay thế CAPM truyền thống. Kết quả kiểm định Wilcoxon cho thấy mô hình SVR dự báo tốt hơn mô hình CAPM truyền thống với giá trị p-value nhỏ hơn 0.05. Một số nhân tố giải thích cho sự biến động của RMSESVR, bao gồm RMSECAPM, VAR, SD, và MEAN; trong đó RMSECAPM là nhân tố có ảnh hưởng lớn nhất, hàm ý rằng sai số trong dự báo của SVR phụ thuộc phần lớn vào mô hình CAPM.

Với những kết quả thu được từ nghiên cứu này, nghiên cứu khuyến nghị các nội dung sau:

- Đối với các nhà đầu tư: nên xem xét mô hình kết hợp SVR thay thế mô hình CAPM truyền thống vì mô hình kết hợp có độ chính xác cao hơn.

- Đối với các nhà nghiên cứu: Các thuật toán máy học khai thác hiệu quả mối quan hệ phức tạp giữa các biến so với các mô hình thống kê kinh tế lượng truyền thống. Do đó, sự kết hợp giữa mô hình lý thuyết và thuật toán máy học kỳ vọng sẽ tạo ra một cuộc cách mạng trong lĩnh vực công nghệ tài chính (FINTECH).

Trong nghiên cứu này, mặc dù mô hình kết hợp giữa SVR và CAPM cho hiệu quả dự báo tốt hơn so với mô hình CAPM đơn lẻ, tuy nhiên sai số dự báo vẫn còn ở mức cao. Do đó, các nghiên cứu tiếp theo có thể tiếp cận thông quá một số thuật toán học máy như mạng nơ-ron hồi quy (Recurrent Neural Network - RNN), mạng nơ-ron nhân tạo (Artificial Neural Network - ANN), mang nơ-ron tích chập (Convolutional Neural Network - CNN), … nhằm cải thiện độ chính xác của dự báo. Ngoài ra, nghiên cứu chỉ xem xét trong bối cảnh ở HOSE, cần mở rộng sang nhiều thị trường tài chính khác nhằm tăng độ tin cậy cho nghiên cứu. Lý thuyết CAPM cho đến nay vẫn còn gây tranh cải do có quá nhiều giả định khó có thể được đáp ứng, do đó, cần thay thế bởi mô hình lý thuyết tốt hơn như mô hình 05 nhân tố.

Tài liệu tham khảo

Abraham, A., Nath, B., & Mahanti, P. K. (2001). *Hybrid intelligent systems for stock market analysis.* Paper presented at the International Conference on Computational Science, San Francisco, California, USA.

Amihud, Y., Christensen, B. J., & Mendelson, H. (1992). *Further evidence on the risk-return relationship* (Vol. 11). Stanford, CA: Stanford University.

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques - Part II: Soft computing methods. *Expert Systems with Applications, 36*(3), 5932-5941.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics, 9*(1), 3-18.

Basu, S. (1983). The relationship between earnings' yield, market value and return for NYSE common stocks: Further evidence. *Journal of Financial Economics, 12*(1), 129-156.

Black, F. (1972). Capital market equilibrium with restricted borrowing. *The Journal of Business, 45*(3), 444-455.

Breen, W. J., & Korajczyk, R. A. (1993). *On selection biases in book-to-market based tests of asset pricing models*. Evanston, IL: Northwestern University.

Bui, K. T., & Thai, T. D. (2021). Capital structure and trade-off theory: Evidence from Vietnam. *The Journal of Asian Finance, Economics, and Business, 8*(1), 45-52. doi:10.13106/jafeb.2021.vol8.no1.045

Bui, K. T., & Tran, H. T. (2021). *Support vector regression algorithm under in the CAPM Framework.* Paper presented at the 2021 International Conference on Data Analytics for Business and Industry (ICDABI), Sakheer, Bahrain. doi:10.1109/ICDABI53623.2021.9655797

Cao, L.-J., & Tay, F. E. H. (2003). Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Transactions on Neural Networks, 14*(6), 1506-1518.

Chaudhary, P. (2017). Testing of CAPM in Indian context. *Business Analyst, 37*(1), 1-18.

Chen, H., Xiao, K., Sun, J., & Wu, S. (2017). A double-layer neural network framework for high-frequency forecasting. *ACM Transactions on Management Information Systems (TMIS), 7*(4), 1-17.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

da Silva Fonseca, J. G., Jr., Oozeki, T., Takashima, T., Koshimizu, G., Uchida, Y., & Ogimoto, K. (2012). Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu, Japan. *Progress in Photovoltaics: Research and Applications, 20*(7), 874-882.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25*(2), 383-417.

Fama, E. F. (1991). Efficient capital markets II. *The Journal of Finance, 46*(5), 1575-1617.

Fama, E. F., & French, K. R. (2021). *The cross-section of expected stock returns*. Chicago, IL: University of Chicago Press.

Fama, E. F., & James, D. (1973). Equilibrium: Empirical tests. *The Journal of Political Economy, 81*(3), 607-636.

Gogas, P., Papadimitriou, T., & Karagkiozis, D. (2018). *The Fama 3 and Fama 5 factor models under a machine learning framework.* Truy cập ngày 10/10/2021 tại https://ideas.repec.org/p/rim/rimwps/18-05.html

Graham, J. R., & Harvey, C. R. (2001). The theory and practice of corporate finance: Evidence from the field. *Journal of Financial Economics, 60*(2/3), 187-243.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018). Stock price prediction using support vector regression on daily and up to the minute prices. *The Journal of Finance and Data Science, 4*(3), 183-201.

Jagannathan, R., & McGrattan, E. R. (1995). The CAPM debate. *Federal Reserve Bank of Minneapolis Quarterly Review, 19*(4), 2-17.

Kumar, M., & Thenmozhi, M. (2014). Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN, and ARIMA-random forest hybrid models. *International Journal of Banking, Accounting and Finance, 5*(3), 284-308.

Lohano, K., & Kashif, M. (2018). Testing asset pricing models on the Pakistan stock exchange. *IBA Business Review, 13*(2), 1-19.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of Economic Perspectives, 17*(1), 59-82.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance, 7*(1), 77-91.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications, 42*(4), 2162-2172.

Qu, H., & Zhang, Y. (2016). A new kernel of support vector regression for forecasting high-frequency stock returns. *Mathematical Problems in Engineering, 2016,* 1-9. doi:10.1155/2016/4907654

Ramedani, Z., Omid, M., Keyhani, A., Shamshirband, S., & Khoshnevisan, B. (2014). Potential of radial basis function based support vector regression for global solar radiation prediction. *Renewable and Sustainable Energy Reviews, 39,* 1005-1011.

Tay, F. E., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega, 29*(4), 309-317.

Tong, J. (2015). The price forecasting of military aircraft based on SVR. *Journal of Computer and Communications, 3*(5), 234-238.

Tran, H. T. (2020). *Application of Machine Learning in CAPM* (Master's thesis, University of Economics Ho Chi Minh City, Ho Chi Minh City, Vietnam). Truy cập ngày 10/10/2021 tại https://opac.ueh.edu.vn/record=b1032827~S8

Tran, K. T., Banh, T. T., & Nguyen, A. H. T. (2012). Dự đoán giá cổ phiếu trên thị trường chứng khoán Việt Nam bằng phương pháp lai GA-SVR [Predicting stock prices on Vietnam stock market by hybrid method GA-SVR]. *Chuyên san Các công trình nghiên cứu, phát triển và ứng dụng Công nghệ thông tin và Truyền thông, V-1*(7), 12-22.

Treynor, J. L. (1961). *Market value, time, and risk.* Truy cập ngày 10/10/2021 tại https://ssrn.com/abstract=2600356

Trinh, N. T. (2013). *Dự đoán xu hướng thị trường chứng khoán bằng cách sử dụng Twitter [Predict stock market trends using Twitter]* (Master's thesis). Vietnam National University Hanoi, Hanoi, Vietnam.

Vapnik, V. (2013). *The nature of statistical learning theory*. New York, NY: Springer Science & Business Media.

Wang, J.-J., Wang, J.-Z., Zhang, Z.-G., & Guo, S.-P. (2012). Stock index forecasting based on a hybrid model. *Omega, 40*(6), 758-766.

Weng, B., Ahmed, M. A., & Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications, 79,* 153-163.

Yuan, F.-C., Lee, C.-H., & Chiu, C. (2020). Using market sentiment analysis and genetic algorithm-based least squares support vector regression to predict gold prices. *International Journal of Computational Intelligence Systems, 13*(1), 234-246.

Zhang, N., Lin, A., & Shang, P. (2017). Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting. *Physica A: Statistical Mechanics and Its Applications, 477,* 161-173.

Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications, 67,* 126-139.