# LỰA CHỌN ĐẶC TRƯNG VÀ DỰ BÁO RỦI RO VỠ NỢ DOANH NGHIỆP: THỰC NGHIỆM VỚI MÔ HÌNH HỌC MÁY

**Nguyễn Minh Nhật[1*], Ngô Hoàng Khánh Duy[1]**

[1]Trường Đại học Ngân hàng Thành phố Hồ Chí Minh
*Tác giả liên hệ: Email: nhatnm@hub.edu.vn

*Ngày nhận:* 16/02/2025 *Ngày nhận lại:* 26/03/2025 *Ngày đăng:* 25/06/2025
DOI: 10.52932/jfmr.v16i3.761

**Phụ lục 1.** Mô tả thông tin các biến trong bộ dữ liệu

| Thông tin các biến | Số lượng | Nhóm các biến | Chi tiết các biến trong nhóm |
|---|---|---|---|
| 95 Biến giải thích được phân chia thành 11 nhóm biến | **14** | Nhóm chỉ số sinh lời | After Tax Net Interest Rate, Continuous Interest Rate After Tax, Gross Profit To Sales, Net Income To Stockholders Equity, Net Income To Total Assets, Net Profit Before Tax To Capital, Operating Gross Margin, Operating Profit Rate, Operating Profit To Capital, Pre Tax Net Interest Rate, Realized Sales Gross Margin, Return On Assets After Tax, Return On Assets After Tax Depreciation, Return On Assets Before Interest Depreciation |
| | **06** | Nhóm chỉ số thanh khoản | Cash To Current Liability, Current Liability To Current Assets, Current Ratio, No Credit Interval, Quick Assets To Current Liability, Quick Ratio |
| | **12** | Nhóm chỉ số đòn bẩy tài chính | Borrowing Dependency, Contingent Liabilities Equity Ratio, Current Liabilities To Equity, Current Liability To Equity, Debt To Asset Ratio, Debt To Equity Ratio, Degree Of Financial Leverage, Equity To Asset Ratio, Equity To Liability, Equity To LongTerm Liability, Liability To Equity, LongTerm Liability To Current Assets |
| | **12** | Nhóm chỉ số hiệu suất hoạt động | Accounts Receivable Turnover, Average Collection Days, Cash Turnover Rate, Current Asset Turnover Rate, Equity Turnover Rate, Fixed Assets Turnover Frequency, Inventory Turnover Rate, Operating Profit Per Person, Quick Asset Turnover Rate, Revenue Per Person, Total Asset Turnover, Working Capital Turnover Rate |
| | **08** | Nhóm chỉ số tăng trưởng | After Tax Net Profit Growth Rate, Continuous Net Profit Growth Rate, Net Value Growth Rate, Operating Profit Growth Rate, Realized Sales Gross Profit Growth Rate, Regular Net Profit Growth Rate, Total Asset Growth Rate, Total Asset Return Growth Rate |
| | **07** | Nhóm chỉ số giá trị cổ phiếu | Adjusted Book Value Per Share, Book Value Per Share, Final Book Value Per Share, Net Profit Before Tax Per Share, Operating Profit Per Share, Persistent Earnings Per Share, Revenue Per Share |
| | **08** | Nhóm chỉ số dòng tiền | Cash Flow From Operations To Assets, Cash Flow Per Share, Cash Flow Rate, Cash Flow To Equity, Cash Flow To Liability, Cash Flow To Sales, Cash Flow To Total Assets, Cash Reinvestment Ratio |

| | | | |
|---|---|---|---|
| | **15** | Nhóm chỉ số cơ cấu tài sản và nguồn vốn | Cash To Total Assets, Current Assets To Total Assets, Current Liabilities To Total Liability, Current Liability To Assets, Current Liability To Total Liability, Fixed Assets To Assets Ratio, Inventory And Receivables To Equity, Inventory To Current Liability, Inventory To Working Capital, Long Term Fund Suitability Ratio, Operating Funds To Liability, Quick Assets To Total Assets, Retained Earnings To Total Assets, Working Capital To Equity, Working Capital To Total Assets |
| | **06** | Nhóm chỉ số chi phí và hiệu quả hoạt động | Allocation Rate Per Person, Non Operating Income Expenditure Ratio, Operating Expense Rate, Research And Development Expense Rate, Total Expense To Assets Ratio, Total Income To Expense Ratio |
| | **04** | Nhóm chỉ số chi phí tài chính và thuế | Effective Tax Rate, Interest Coverage Ratio, Interest Expense Ratio, InterestBearing Debt Interest Rate |
| | **03** | Nhóm chỉ số khác | Liability Assets Flag, Net Income Flag, Total Assets To GNP Price |
| Biến mục tiêu | 01 | | "Bankrupt?" |

*Nguồn: Bộ dữ liệu Kaggle (2024)*

**Phụ lục 2.** Phương pháp lựa chọn đặc trưng

*(1) Loại bỏ đặc trưng đệ quy (Recursive Feature Elimination - RFE)*

RFE là một phương pháp dạng bọc (wrapper method) tập trung vào việc giảm tầm quan trọng của các đặc trưng không liên quan, vốn có thể làm suy giảm hiệu suất của mô hình. Một ưu điểm nổi bật của RFE là khả năng xử lý hiệu quả vấn đề nhiễu trong phân tích các tập dữ liệu phức tạp, nhờ đó các mô hình dự báo có thể đạt được hiệu suất tối ưu. RFE không chỉ đảm bảo tính hiệu quả của mô hình dự đoán mà còn giữ lại các đặc trưng thông tin nhất, giúp chúng dễ dàng được diễn giải. Tuy nhiên, một nhược điểm lớn của RFE là yêu cầu cao về tính toán, vì cần huấn luyện mô hình nhiều lần để đánh giá từng đặc trưng, dẫn đến tiêu tốn nhiều thời gian và tài nguyên. Trong nghiên cứu này, kỳ vọng rằng RFE sẽ đạt được hiệu suất cao nhất, đặc biệt khi được kết hợp với LightGBM – mô hình có khả năng khai thác tối đa các đặc trưng phi tuyến (Li và cộng sự, 2017).

Trong mỗi vòng lặp, RFE huấn luyện một mô hình (kí hiệu M) trên bộ dữ liệu sử dụng tập đặc trưng hiện tại $F_t$. Sau đó, độ quan trọng $\mathrm{Imp}_t(j)$ của từng đặc trưng j trong $F_t$ được ước lượng để tìm kiếm đặc trưng kém quan trọng nhất, nhằm loại bỏ nó khỏi quá trình huấn luyện ở vòng tiếp theo. Quá trình này lặp lại cho đến khi số lượng đặc trưng còn lại đạt đúng ngưỡng kỳ vọng. RFE có thể được mô tả dưa trên công thức giả lập (Pseudo-formula) như sau:

*Khởi tạo:* $F_0 = \{1,2,\dots,m\}, t = 0.$

*Trong khi* $|F_t| > k$*:*

1. *Huấn luyện mô hình M với các đặc trưng thuộc* $F_t$.
2. *Tính* $Imp_t(j)$ *cho mỗi* $j \in F_t$
3. $j_{min} = arg\min_{j \in F_t} Imp_t(j)$
4. $F_{t+1} = F_t \setminus \{j_{min}\}$
5. $t \leftarrow t + 1$

*Kết thúc khi* $|F_t| = k$

*Đầu ra cuối cùng: $F_t$ (tập đặc trưng được giữ lại).*

*(2) Loại bỏ đặc trưng chuyển tiếp (Forward Feature Selection - FFS)*

FFS là một quy trình thêm các đặc trưng lần lượt theo mức độ quan trọng đối với hiệu suất của mô hình. Mặc dù thuật toán này khá đơn giản, trực quan và yêu cầu ít tài nguyên hơn, nhưng nó lại tương thích với hầu hết các thuật toán học máy (Xia & Yang, 2022). Ngoài ra, cách tiếp cận theo FFS cũng có thể được các nhà đầu tư sử dụng thành công để khám phá mối quan hệ giữa các biến độc lập và biến phụ thuộc trong một lĩnh vực tài chính cụ thể. Tuy phương pháp FFS dễ triển khai và phù hợp với nhiều ứng dụng thực tiễn, tuy nhiên cách thức tiếp cận của FFS lại dễ rơi vào cực tiểu cục bộ, khi LCĐT từng bước một và có thể bỏ qua những tổ hợp đặc trưng ở các bước sau. Bên cạnh đó, FFS cũng không xử lý tốt mối quan hệ phi tuyến giữa các đặc trưng trong mô hình (Fallahpour & Piri, 2017).

Quy trình bắt đầu từ một tập đặc trưng rỗng ($\emptyset$), sau đó, trong mỗi vòng lặp, mô hình được huấn luyện lần lượt với từng đặc trưng tiềm năng để xác định xem đặc trưng nào mang lại mức cải thiện hiệu suất cao nhất khi kết hợp với tập đặc trưng hiện tại. Nhờ cách thức chọn lựa gia tăng, FFS thường được coi là trực quan, dễ triển khai, đồng thời có thể tiết kiệm tài nguyên hơn so với một số phương pháp phức tạp khác. FFS có thể được mô tả dựa trên công thức giả lập (Pseudo-formula) như sau:

*Khởi tạo:* $S_0 = \emptyset, t = 0$.

*Trong khi* $|S_t| < k$ *(hoặc còn đặc trưng chưa được chọn):*

1. *Với mỗi đặc trưng* $f \in F \setminus S_t$*, huấn luyện mô hình M sử dụng tập* $S_t \cup \{f\}$
2. *Xác định* $f^* = \underset{f \in F \setminus S_t}{arg\,max}\, \Delta$ *(hiệu suất)*
3. $S_{t+1} = S_t \cup \{f^*\}$
4. $t \leftarrow t + 1$

*Kết thúc khi* $S_t$ *đã đạt số lượng hoặc tiêu chí dừng*

*Đầu ra:* $S_t$ *(tập đặc trưng được chọn)*

*(3) Loại bỏ đặc trưng ngược (Backward Feature Elimination - BFE)*

BFE bắt đầu với toàn bộ tập hợp các đặc trưng và loại bỏ lần lượt những đặc trưng ít giá trị nhất. Kỹ thuật này đặc biệt hữu ích khi được áp dụng trên các tập dữ liệu có kích thước lớn bằng cách giảm bớt chiều dữ liệu và cải thiện khả năng tổng quát hóa của mô hình. Bằng việc loại bỏ đều đặn các đặc trưng kém đóng góp, BFE hỗ trợ những mô hình dự đoán—bao gồm cả các mô hình tuyến tính và phi tuyến—tập trung hơn vào các đặc trưng giàu thông tin, đồng thời giảm rủi ro quá khớp (overfitting) của mô hình. Dù yêu cầu nhiều tài nguyên tính toán do cần huấn luyện và đánh giá mô hình nhiều lần, BFE vẫn được đánh giá cao bởi khả năng cung cấp một tập đặc trưng vừa gọn nhẹ vừa có sức giải thích tốt, đặc biệt khi được áp dụng trong các môi trường dữ liệu phức tạp (Fallahpour & Piri, 2017). BFE có thể được mô tả dựa trên công thức giả lập (Pseudo-formula) như sau:

*Khởi tạo:* $S_0 = \{1,2,\dots,m\}, t = 0$.

*Trong khi* $|S_t| > k$*:*

1. *Huấn luyện mô hình M với các đặc trưng thuộc* $S_t$.
2. *Tính* $Imp_t(j)$ *cho mỗi* $j \in S_t$
3. $j_{min} = \underset{j \in S_t}{arg\,min}\, Imp_t(j)$
4. $S_{t+1} = S_t \setminus \{j_{min}\}$
5. $t \leftarrow t + 1$

*Kết thúc khi* $|S_t| = k$

*Đầu ra:* $S_t$ *(tập đặc trưng được giữ lại).*

Trong đó, $\mathrm{Imp}_t(j)$ phản ánh mức độ quan trọng của đặc trưng j tại vòng lặp t, có thể được tính dựa trên nhiều chỉ báo khác nhau, từ trị tuyệt đối của hệ số hồi quy đến mức giảm impurity (đối với mô hình cây). Ở mỗi bước, đặc trưng kém quan trọng nhất $j_{min}$ sẽ được loại bỏ, nhờ đó mô hình dần tập trung vào những thuộc tính cốt lõi. Tuy chi phí tính toán có thể tăng do yêu cầu huấn luyện lặp, BFE mang lại một tập đặc trưng tinh gọn, vừa nâng cao độ chính xác của mô hình, vừa duy trì khả năng diễn giải cao trong các bối cảnh ứng dụng khác nhau.

*(4) Loại bỏ đặc trưng dựa trên hoán vị (Permutation Feature Importance - PFI)*

PFI dựa trên việc giảm hiệu suất của mô hình khi giá trị của một đặc trưng được xáo trộn ngẫu nhiên. Với cách tiếp cận này, PFI cung cấp một thước đo tổng quát về tầm quan trọng của từng đặc trưng trong việc dự đoán RRVN. Bên cạnh đó, PFI cũng không yêu cầu huấn luyện lại mô hình sau mỗi lần hoán vị, điều này giúp cho phương pháp tiết kiệm được thời gian so với phương pháp khác. Tuy nhiên, PFI cũng có một số hạn chế nhất định như khá là nhạy cảm với dữ liệu, đặc biệt khi dữ liệu mất cân bằng hoặc chứa các giá trị ngoại lai. Ngoài ra, PFI cũng không hoạt động tốt cho dữ liệu chuỗi thời gian hoặc dữ liệu có mức độ tương quan lớn (Gregorutt và cộng sự, 2010). PFI có thể được mô tả dựa trên công thức giả lập (Pseudo-formula) như sau:

*Khởi tạo:*

1. *Xác định hiệu suất ban đầu của mô hình M trên tập đặc trưng đầy đủ* $S_0$*, ký hiệu là* $perf^{orig}$.
2. *Thiết lập* $S_0 = F$ *(tập đặc trưng đầy đủ).*

*Vòng lặp đánh giá đặc trưng*

- *Với mỗi đặc trưng* $f \in S_t$*:*
  1. *Xáo trộn (hoán vị) ngẫu nhiên giá trị của f để tạo ra tập dữ liệu mới* $X_f^{perm}$ .
  2. *Huấn luyện mô hình M trên* $X_f^{perm}$ *và đánh giá hiệu suất* $Perf_f^{perm}$.
  3. *Xác định mức độ quan trọng của f theo công thức:*

$$Imp_t(f) = perf^{orig} - Perf_f^{perm}$$

- *Xác định đặc trưng ít quan trọng nhất:*

$$f^{min} = arg\min_{f \in S_t} Imp_t(f)$$

- *Loại bỏ đặc trưng ít quan trọng nhất:*

$$S_{t+1} = S_t \setminus \{f_{min}\}$$

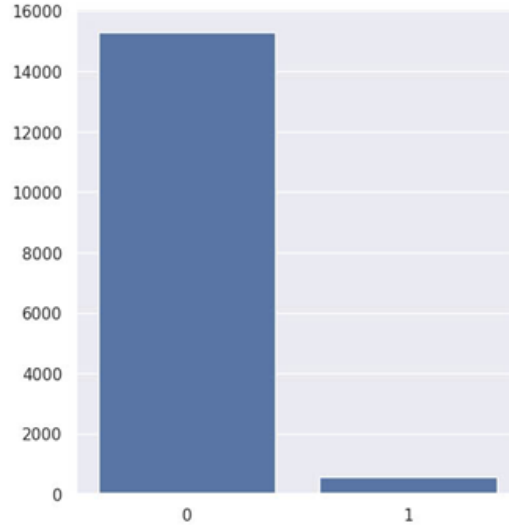
- *Cập nhật* $t \leftarrow t + 1$.

*Kết thúc khi* $S_t$ *đạt số lượng đặc trưng mong muốn hoặc tiêu chí dừng.*

*Đầu ra: Tập đặc trưng được chọn* $S_t$.

Trong đó, mức độ quan trọng $Imp_t(f)$ phản ánh sự suy giảm hiệu suất mô hình khi hoán vị đặc trưng f tại vòng lặp t, được xác định bằng cách so sánh hiệu suất của mô hình trước và sau khi hoán vị ngẫu nhiên giá trị của đặc trưng đó. Mức độ quan trọng của đặc trưng f được tính bằng chênh lệch giữa hiệu suất ban đầu và hiệu suất sau hoán vị, theo công thức $Imp_t(f) = perf^{orig} - Perf_f^{perm}$. Đặc trưng có mức độ quan trọng thấp nhất $f^{min}$ được xác định và loại bỏ khỏi tập đặc trưng $S_t$, nhờ đó mô hình dần tập trung vào các đặc trưng quan trọng hơn. Quy trình tiếp tục cho đến khi đạt được số lượng đặc trưng mong muốn hoặc thỏa mãn tiêu chí dừng.

**Phụ lục 3.** Mô tả về biến mục tiêu




| Tình trạng vỡ nợ | Số lượng | Tỷ lệ |
|---|---|---|
| 0 | 15,284 | 96,36% |
| 1 | 578 | 3,64% |

**Phụ lục 4.** LCĐT quan trọng của 04 phương pháp RFE, FFS, BFE, và PFI