# CÁC MÔ HÌNH MÁY HỌC PHÙ HỢP CHO DỰ BÁO THU NGÂN SÁCH NHÀ NƯỚC CỦA VIỆT NAM

**Ths. Tạ Văn Thắng* - Ths. Phùng Thị Khang Ninh** - Tạ Thị Trang*****

***Bài báo đánh giá hiệu quả ứng dụng các mô hình học máy trong dự báo thu ngân sách nhà nước (NSNN) của Việt Nam trên dữ liệu giai đoạn 1991-2024, qua đó khẳng định giá trị gia tăng rõ rệt của học máy trong việc mô hình hóa các quan hệ phi tuyến và phản ứng linh hoạt trước các cú sốc chính sách tài khóa. Kết quả thực nghiệm cho thấy không tồn tại "một mô hình tốt cho mọi khoản thu", song cách tiếp cận theo danh mục mô hình tối ưu hóa riêng cho từng khoản thu giúp nâng cao độ chính xác tổng thể và giảm sai số dự báo tổng thu xuống chỉ còn 1-2% mỗi năm. Cấu hình khuyến nghị gồm: RFcho các sắc thu nội địa quy mô lớn (PIT, VAT, EXT), ElasticNet cho các khoản thu có cấu trúc gần tuyến tính và dữ liệu hạn chế (CIT, EMT, AID), XGBoost cho nhóm thu chịu biến động theo chu kỳ bất động sản hoặc cú sốc chính sách (LAND, LP), và OLS cho các khoản thu ổn định, có quan hệ tuyến tính rõ (OIL, OTH). Trong giai đoạn tới, nghiên cứu đề xuất triển khai mô hình lai phần dư hai giai đoạn (two-stage residual modeling) để kết hợp ưu thế giải thích của mô hình kinh tế lượng và khả năng học phi tuyến của mô hình máy học, hướng tới hệ thống dự báo tài khóa chủ động, cập nhật theo thời gian thực trong bối cảnh chuyển đổi số tài chính công.***

• Từ khóa: *dự báo thu NSNN; học máy; random forest; XGBoost; ElasticNet; mô hình lai .*

*This study evaluates the performance of machine learning models in forecasting Vietnam's state budget revenue (SBR) using annual data from 1991–2024, highlighting the significant added value of machine learning in capturing nonlinear relationships and adapting to fiscal policy shocks. The empirical results reveal that there is no "one-size-fits-all" model for all revenue categories; however, adopting a portfolio approach with model optimization by revenue component substantially improves overall accuracy, reducing total forecast error to just 1-2% per year. The recommended configuration includes: Random Forest for major domestic tax revenues (PIT, VAT, and excise taxes), ElasticNet for revenue items with near-linear structures and limited data (CIT, EMT, AID), XGBoost for revenues influenced by real estate cycles or policy shocks (LAND, LP), and OLS for stable, linearly related items (OIL, OTH). For future research, a two-stage residual hybrid modeling framework is proposed to combine the interpretability of econometric models with the nonlinear learning capacity of machine learning algorithms paving the way for a real-time, adaptive fiscal forecasting system within Vietnam's digital public finance transformation.*

• Key words: *state budget revenue forecasting; machine learning; Random Forest; XGBoost; ElasticNet; hybrid modeling.*

*Ngày nhận bài: 20/9/2025*
*Ngày gửi phản biện: 21/10/2025*
*Ngày duyệt đăng: 21/11/2025*
*DOI: https://doi.org/10.71374/jfarv.v25.i302.04*

## 1. Bối cảnh và yêu cầu đổi mới phương pháp dự báo

Khác với các nền kinh tế phát triển có hệ thống dữ liệu tài khóa tập trung, thống nhất và được chuẩn hóa cao, Việt Nam vẫn đang trong giai đoạn đầu của quá trình hiện đại hóa hệ thống thống kê và quản lý tài chính công. Cấu trúc thu NSNN của Việt Nam chịu tác động đồng thời của các thay đổi chính sách nội sinh, cú sốc bên ngoài và sự chuyển đổi mạnh mẽ của nền kinh tế.

Trong hơn ba thập kỷ qua, Việt Nam đã trải qua nhiều điểm gãy cấu trúc trong chuỗi dữ liệu thu NSNN: (i) Thay đổi hệ thống thuế cơ bản: các cải cách lớn về thuế CIT và VAT (1997, 2008, 2014, 2020, 2025 - Luật số 48/2024/QH15) đã liên tục điều chỉnh quan hệ giữa thuế và các biến vĩ mô như GDP, tiêu dùng, đầu tư. Việc giảm thuế suất CIT từ 25% xuống 20% và giảm VAT từ 10% xuống 8% (2022-2026) tạo cú sốc chính sách rõ rệt; (ii) Bổ sung các nguồn thu mới: Luật Thuế Bảo vệ môi trường (2010, hiệu lực 2012), cùng với việc chính thức đưa thu xổ số kiến thiết, phí - lệ phí, viện trợ và dầu khí vào NSNN, làm thay đổi cơ cấu và mức độ biến động của chuỗi thu; (iii) Cú sốc bên ngoài và hội nhập quốc tế: việc Việt Nam gia nhập WTO (2007), thực thi ATIGA, CPTPP, EVFTA và RCEP khiến các khoản thu từ XNK giảm mạnh và thể hiện tính chu kỳ, phi tuyến cao; (iv) Cú sốc nội sinh: đại dịch COVID-19 (2020-2021) cùng chính sách giãn, hoãn và giảm thuế làm đứt gãy chuỗi thu - đặc biệt ở VAT, CIT và PIT; (v) Đánh giá lại quy mô GDP (2018-2019) theo chuẩn SNA 2008 khiến

** Viện Chiến lược và Chính sách kinh tế - tài chính, Bộ Tài chính; email: tavanthang@mof.gov.vn*
*** Trường Đại học Hùng Vương*
**** Thuế cơ sở 24 Thành phố Hà Nội*

GDP danh nghĩa gia tăng, phá vỡ tính ổn định của các quan hệ kinh tế.

Những đặc điểm này khiến các mô hình truyền thống như hồi quy tuyến tính (OLS), hệ phương trình đồng thời (SUR) hay các mô hình chuỗi thời gian (ARIMA, VECM) từng giữ vai trò trung tâm trong dự báo thu NSNN trở nên hạn chế: chúng chỉ mô tả được xu hướng trung bình nhưng không phản ứng nhanh với biến động ngắn hạn hay thay đổi thể chế. Trong bối cảnh đó, việc tích hợp các mô hình học máy trong dự báo thu NSNN Việt Nam là một bước chuyển từ dự báo tĩnh sang hoạch định tài khóa chủ động, giúp nâng cao độ chính xác, khả năng thích ứng và hỗ trợ điều hành ngân sách theo thời gian thực - phù hợp với tiến trình chuyển đổi số trong tài chính công hiện nay.

**2. Các mô hình máy học tiềm năng trong dự báo thu Ngân sách nhà nước**

Sự phát triển của khoa học dữ liệu và năng lực tính toán đã mở ra khả năng ứng dụng mô hình học máy (machine learning) trong lĩnh vực dự báo tài khóa. Không chỉ dừng ở việc thay thế công cụ, học máy cho phép mô hình "học" từ dữ liệu - phát hiện tự động các mẫu ẩn, xu hướng phi tuyến và tương tác đa chiều giữa các yếu tố kinh tế, chính sách và hành vi thị trường. Một số mô hình có tiềm năng ứng dụng hiệu quả trong dự báo thu NSNN có thể kể đến:

***2.1. Nhóm mô hình cây quyết định: Random Forest và Extreme Gradient Boosting***

Hai mô hình này được xây dựng trên nền tảng cây quyết định (decision tree) - một thuật toán phân tách dữ liệu thành các nhánh nhỏ dựa trên giá trị của biến giải thích, nhằm tối thiểu hóa sai số dự báo trong từng bước chia.

Mô hình Random Forest (RF): là sự mở rộng trực tiếp của cây quyết định, trong đó hàng trăm hoặc hàng nghìn cây được huấn luyện trên các tập mẫu và tập biến ngẫu nhiên khác nhau, sau đó kết quả dự báo cuối cùng được lấy bằng trung bình (hoặc biểu quyết) của toàn bộ rừng cây. Cơ chế này giúp giảm phương sai và hạn chế hiện tượng quá khớp - vốn thường gặp trong các mô hình tuyến tính hoặc cây đơn. Mỗi cây chỉ học một phần nhỏ cấu trúc của dữ liệu, nên tổng hợp của nhiều cây mang lại bức tranh ổn định hơn về quan hệ phi tuyến.

RF phù hợp nhất với các khoản thu có nhiều yếu tố chi phối, quan hệ phi tuyến vừa phải, và sự thay đổi chính sách lặp lại - ví dụ thuế thu nhập doanh nghiệp, thuế VAT nội địa, hoặc thu xuất nhập khẩu.

Extreme Gradient Boosting (XGBoost): Trong khi RFhọc song song bằng cách trung bình nhiều cây độc lập, XGBoost lại học tuần tự (boosting) - nghĩa là mỗi cây mới được xây dựng để sửa sai cho cây trước đó. Cụ thể, XGBoost tối ưu hàm mất mát bằng thuật toán gradient descent, với cơ chế trọng số giúp cây sau "tập trung" vào các quan sát mà mô hình trước dự báo sai. Kết quả là, XGBoost thường đạt độ chính xác cao hơn trong các chuỗi dữ liệu có tương tác phức tạp giữa nhiều biến vĩ mô và chính sách.

Một cải tiến đáng kể của XGBoost so với mô hình Boosting truyền thống là nó tự động điều chỉnh để tránh quá khớp, đồng thời tối ưu hóa tốc độ tính toán thông qua nén dữ liệu và xử lý song song. Do đó, mô hình có thể huấn luyện nhanh và hiệu quả ngay cả khi có hàng chục biến đầu vào và số quan sát dài.

Ưu thế lớn nhất của nhóm mô hình cây quyết định là khả năng xử lý dữ liệu phi tuyến, nhiều chiều và có biến định tính, vốn là đặc trưng của hệ thống thu Việt Nam. RF cung cấp dự báo ổn định và dễ kiểm soát, trong khi XGBoost mang lại độ chính xác cao hơn nhờ học sâu hơn từ sai số. Ngoài ra, hai mô hình này còn cung cấp chỉ số tầm quan trọng của biến (feature importance) - giúp nhận diện các yếu tố vĩ mô hoặc chính sách có ảnh hưởng lớn nhất đến từng khoản thu.

***2.2. Nhóm mô hình dựa trên khoảng cách: K-Nearest Neighbors (KNN)***

Mô hình K-Nearest Neighbors (KNN) là một trong những thuật toán học máy đơn giản nhưng có tính ứng dụng cao trong dự báo chuỗi thời gian tài chính - ngân sách. Nguyên tắc hoạt động của KNN có thể tóm gọn là: *"năm nay giống năm nào nhất trong quá khứ"*. Nghĩa là, thay vì xây dựng một hàm hồi quy cố định giữa biến độc lập và biến phụ thuộc, KNN tìm kiếm k quan sát trong quá khứ có đặc điểm gần nhất với năm cần dự báo, rồi lấy trung bình có trọng số của các quan sát đó để suy ra giá trị tương lai.

Cụ thể, với mỗi năm t, mô hình xác định một vector đặc trưng $X_t = (GDP_t, C_t, I_t,\ldots)$ gồm các biến vĩ mô đầu vào. Khi cần dự báo giá trị thu ngân sách $Y_t$, thuật toán sẽ tính khoảng cách Euclid (hoặc Manhattan,…) giữa Xtvà toàn bộ các vector trong tập dữ liệu quá khứ. Các năm có đặc điểm gần nhất (ví dụ cùng mức tăng GDP, tỷ giá và giá dầu) được chọn làm "hàng xóm gần nhất" và dự báo Ŷt được tính như trung bình có trọng số nghịch đảo với khoảng cách đó theo công thức:

$$\hat{Y}_t = \frac{\sum_{i=1}^{k} \frac{Y_i}{d(X_t, X_i)}}{\sum_{i=1}^{k} \frac{1}{d(X_t, X_i)}}$$

Điểm mạnh lớn nhất của KNN là không cần giả định hàm hồi quy, do đó rất phù hợp với dữ liệu ngắn, không ổn định hoặc có mối quan hệ phi tuyến nhưng khó xác định rõ dạng. Mô hình này cũng tự động thích ứng với các biến mới, ví dụ khi xuất hiện cú sốc chính sách, chỉ cần bổ sung dữ liệu năm đó vào tập huấn luyện, KNN sẽ học lại mà không phải ước lượng lại toàn bộ hàm.

### *2.3. Nhóm mô hình mạng nơ-ron nhân tạo (Neural Networks - ANN, MLP, LSTM, GRU)*

Các mô hình mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN) là đại diện tiêu biểu cho xu hướng học sâu (Deep Learning) trong kinh tế lượng hiện đại. ANN có khả năng xấp xỉ các hàm phi tuyến phức tạp mà không cần xác định trước dạng hàm. Nhờ đó, nhóm mô hình này đặc biệt phù hợp để mô hình hóa các chuỗi dữ liệu có cấu trúc động, có độ trễ, và chịu tác động phi tuyến mạnh.

Một mạng nơ-ron gồm ba tầng cơ bản: lớp đầu vào (input layer), các lớp ẩn (hidden layers), và lớp đầu ra (output layer). Mỗi nơ-ron trong mạng học cách biến đổi dữ liệu thông qua các trọng số (weights) và hàm kích hoạt (activation functions, thường là ReLU, Sigmoid, hoặc Tanh), nhằm giảm sai số dự báo tổng thể.

Mô hình Multi-Layer Perceptron (MLP) - dạng cơ bản của ANN - học mối quan hệ phi tuyến tĩnh, trong khi các mô hình mạng nơ-ron hồi quy (Recurrent Neural Networks - RNN) như LSTM (Long Short-Term Memory) và GRU (Gated Recurrent Unit) lại có khả năng "ghi nhớ" trạng thái trước đó, giúp mô hình hóa các chuỗi thời gian có quan hệ phụ thuộc theo thời gian (temporal dependence) - điều mà các mô hình như OLS hay RF không thể nắm bắt.

### *2.4. Mô hình lai ghép (Hybrid Models)*

Mô hình lai ghép (Hybrid Model) được phát triển nhằm kết hợp các mô hình thống kê kinh tế lượng truyền thống với mô hình học máy hiện đại, tận dụng ưu thế của cả hai:

- Mô hình truyền thống (OLS, ARIMA, VAR, SUR) cung cấp khả năng giải thích kinh tế học rõ ràng và mô tả xu hướng tuyến tính dài hạn.

- Mô hình học máy (RF, XGBoost, LSTM, GRU) có khả năng bắt các quy luật phi tuyến, tính mùa vụ, và hiệu ứng trễ chính sách mà mô hình tuyến tính thường bỏ sót.

Cấu trúc phổ biến của mô hình lai là tách chuỗi dữ liệu dự báo $y_t$ thành hai phần:

$$y_t = \hat{y}_t^{(linear)} + \hat{y}_t^{(nonlinear)} + \varepsilon_t$$

Trong đó:

$\hat{y}_t^{(linear)}$ được ước lượng bằng mô hình tuyến tính (OLS, ARIMA hoặc VAR) - mô tả xu hướng và quan hệ vĩ mô dài hạn;

$\hat{y}_t^{(nonlinear)}$ là phần dư được mô hình hóa bằng học máy (XGBoost, LSTM, hoặc MLP), nhằm "học" các yếu tố phi tuyến, nhiễu ngắn hạn hoặc cú sốc chính sách.

Cách tiếp cận này có thể hình dung như mô hình hóa phần dư hai giai đoạn: mô hình đầu tiên ước lượng phần có thể giải thích bằng lý thuyết kinh tế, còn mô hình thứ hai xử lý phần phi tuyến và biến động khó dự đoán.

## 3. Kết quả ứng dụng các mô hình máy học trong dự báo thu ngân sách nhà nước

### *3.1. Thiết kế thực nghiệm và tiêu chí đánh giá*

Nghiên cứu thực hiện trên tập dữ liệu theo năm giai đoạn 1991-2024 với 11 khoản thu chính trong cơ cấu thu ngân sách. Bộ biến đầu vào gồm 13 chỉ tiêu vĩ mô chủ chốt (GDP, đầu tư, tiêu dùng chính phủ, tiêu dùng tư nhân, FDI, xuất khẩu, nhập khẩu, tỷ giá, giá dầu) cùng các biến trễ và biến giả chính sách (COVID-19, cải cách CIT, cắt giảm VAT, v.v.).

Nghiên cứu sử dụng chiến lược dự báo mở rộng từng bước, tức là, mỗi năm trong giai đoạn nghiên cứu lần lượt đóng vai trò quan sát ngoài mẫu, được dự báo bởi mô hình huấn luyện trên toàn bộ dữ liệu của các năm trước đó. Các mô hình sẽ được so sánh, đánh giá dựa trên các thước đo RMSE, MAE và $R^2$, mô hình tốt nhất sẽ được lựa chọn ưu tiên RMSE nhỏ nhất.

### *3.2. Kết quả: mô hình tối ưu và diễn giải*

*Với thuế TNCN (PIT_R), mô hình RF* có RMSE thấp nhất và $R^2 \approx 0{,}98$. Tập biến quan trọng gồm các trễ của đầu tư (I_R_lag1, I_R_lag2), GDP_R, IP_R và P_oil cho thấy biến động của thu nhập chịu tác động tổng hợp của chu kỳ đầu tư và tăng trưởng, trong khi giá dầu đóng vai trò biến điều kiện (phản ánh chu kỳ hàng hóa/tỷ giá). RF xử lý tốt tương tác và nhiễu, vượt OLS rõ rệt.

*Với thuế TNDN (CIT_R), mô hình ElasticNet có* RMSE thấp nhất và $R^2 \approx 0{,}95$, tập biến quan trọng gồm CG_R, IG_R, FDI_R, các trễ và P_oil. ElasticNet nhấn mạnh cấu phần tuyến tính ổn định giữa lợi nhuận doanh nghiệp và cầu đầu tư/công.

*Với thuế GTGT hàng nội địa (VAT_R), mô hình tốt nhất là RF với RMSE thấp và $R^2 \approx 0{,}96$.* Biến quan trọng gồm I_R_lag2, P_oil_lag1, COVID, các chỉ tiêu tiêu dùng cho thấy thuế GTGT hàng nội địa chịu tác động lớn từ các chính sách ngắn hạn, RF học tốt các "điểm gãy" (như giai đoạn giảm VAT 8%) mà không cần quy định trước dạng hàm.

*Thuế TTĐB hàng nội địa (EXT_R), RF có RMSE thấp nhất và $R^2 \approx 0{,}87$,* các biến quan trọng gồm M_R_lag1, IG_R, các dummy chính sách (VAT_cut, CIT_reform) và P_oil_lag2. Khoản thu này nhạy với chu kỳ tiêu dùng hàng hóa chịu TTĐB và độ trễ cầu nhập khẩu; RF cân bằng tốt giữa mức độ nhiễu và phi tuyến nhẹ.

*Thu từ đất đai (LAND_R), XGBoost là mô hình có RMSE thấp nhất và $R^2 \approx 0{,}89$.* Các biến quan trọng: CIT_reform, GDP_R, CP_R, IG_R… Dữ liệu LAND_R có cú sốc lớn theo chu kỳ bất động sản (2007-2011 bùng nổ; 2023-2024 đóng băng). XGBoost với boosting tuần tự mô hình hóa tốt các

pha lên/xuống, nhỉnh hơn RF/OLS.

**Bảng: Các mô hình tối ưu lựa chọn cho từng khoản thu NSNN**

| Các khoản thu thành phần | Họ mô hình | RMSE (nghìn tỷ đồng) | MAE (nghìn tỷ đồng) | R-Square | Các nhân tố chính (biến độc lập) |
|---|---|---|---|---|---|
| Thu từ viện trợ | ElasticNet | 3,1524 | 2,3482 | 0,0161 | GDP_R_lag2, X_R, VAT_cut, COVID, P_oil_lag1, C_R_lag2, M_R, P_oil, FDI_R_lag1, IG_R |
| Thuế thu nhập doanh nghiệp (không kể dầu thô) | ElasticNet | 10,6760 | 8,6191 | 0,9514 | CG_R, IG_R, P_oil_lag1, FDI_R, P_oil, FDI_R_lag1, P_oil_lag2, X_R_lag1, C_R_lag2, I_R_lag1 |
| Thuế XNK; thuế TTĐB, thuế BVMT hàng nhập khẩu | ElasticNet | 7,6101 | 6,4650 | 0,5863 | I_R_lag1, I_R, P_oil_lag2, CIT_reform, CG_R, GDP_R, CP_R, C_R, X_R, M_R |
| Thuế TTĐB thu từ hàng hóa sản xuất trong nước | RandomForest | 6,2493 | 4,9044 | 0,8681 | M_R_lag1, IG_R, VAT_cut, GDP_R, IP_R, COVID, GDP_R_lag1, I_R_lag2, CIT_reform, P_oil_lag2 |
| Các khoản thu từ đất | XGBoost | 14,6345 | 11,1848 | 0,8872 | CIT_reform, GDP_R, CP_R, C_R, IP_R, FDI_R, IG_R, CG_R, X_R, M_R |
| Các khoản phí và lệ phí | XGBoost | 3,8781 | 3,2557 | 0,9261 | CIT_reform, GDP_R, CP_R, C_R, IP_R, FDI_R, IG_R, CG_R, X_R, M_R |
| Thu từ dầu thô | OLS | 10,8131 | 8,3254 | 0,8960 | P_oil, I_R_lag1, P_oil_lag2, P_oil_lag1, FDI_R_lag2, FDI_R_lag1, I_R_lag2, CIT_reform, FDI_R, X_R_lag2 |
| Thu nội địa khác | OLS | 19,3734 | 12,9745 | 0,9154 | IG_R, P_oil, FDI_R_lag2, FDI_R_lag1, I_R_lag2, COVID, CIT_reform, CP_R, I_R, X_R |
| Thuế thu nhập cá nhân | RandomForest | 4,0875 | 2,8141 | 0,9809 | I_R_lag2, I_R_lag1, P_oil, GDP_R, IP_R, FDI_R, CP_R, C_R, I_R, X_R |
| Thuế giá trị gia tăng hàng nhập khẩu | RandomForest | 14,5762 | 11,4380 | 0,5634 | P_oil, P_oil_lag1, I_R_lag2, I_R, CG_R, IP_R, CP_R, GDP_R, IG_R, X_R |
| Thuế GTGT thu từ hàng hóa sản xuất kinh doanh trong nước | RandomForest | 7,4339 | 5,8363 | 0,9646 | I_R_lag2, P_oil_lag1, COVID, I_R, I_R_lag1, GDP_R, CP_R, C_R, X_R, M_R |

*Thu từ phí và lệ phí (LP_R), XGBoost cũng là mô hình tốt nhất với $R^2 \approx 0,93$.* Chuỗi ổn định, nhưng có tương tác với quy mô hoạt động kinh tế và thể chế về phí. XGBoost giảm thiên lệch (bias) so với RF trong bối cảnh ít nhiễu, cho RMSE thấp nhất.

*Thu từ dầu thô (OIL_R), lại có mô hình tốt nhất là OLS với $R^2 \approx 0,90$,* RMSE thấp nhất. Bộ biến: P_oil, các trễ P_oil, I_R, FDI trễ. Với OIL_R, quan hệ gần tuyến tính (theo log) giữa giá dầu, sản lượng/quy đổi và thu ngân sách khiến OLS đủ mạnh, trong khi mô hình phi tuyến không tạo thêm lợi ích ròng.

*Thu nội địa khác (OTH_R), mô hình được lựa chọn là OLS với $R^2 \approx 0,92$.* Bộ biến: IG_R, P_oil, FDI trễ, I_R_lag2, COVID, CIT_reform cho thấy quan hệ dạng tuyến tính những biến quy mô tổng hợp và yếu tố chính sách lớn quyết định biến động của chuỗi.

*Thu từ hoạt động xuất nhập khẩu (EMT_R), lựa chọn ElasticNet với $R^2 \approx 0,59$.* Biến then chốt: I_R, I_R_lag1, P_oil_lag2, CG_R, GDP_R. Đây là khoản thu chịu tác động đồng thời của thương mại, tỷ giá và chính sách; ElasticNet tận dụng phần tuyến tính trong khi điều chuẩn hạn chế dao động hệ số.

*Thuế GTGT hàng nhập khẩu (VATM_R), RF có $R^2 \approx 0,56$.* Biến nổi bật: P_oil, trễ I_R, IG_R, IP_R, GDP_R. Dù $R^2$ không cao như nhóm nội địa, RF vẫn là mô hình có độ sai lệch thấp nhất nhờ độ bền vững trước các pha giảm thuế quan/hội nhập (ATIGA, CPTPP, EVFTA).

*Thu viện trợ (AID_R)* mang tính rời rạc và chịu điều tiết hành chính/quốc tế. Tất cả các mô hình đều có $R^2$ thấp thậm chí hầu hết có giá trị âm. ElasticNet có RMSE nhỏ nhất trong nhóm và $R^2$ ~0,02 do bảo toàn sự ổn định hệ số. Mặc dù, ElasticNet được chọn nhưng khoản thu này phù hợp cho dự báo bằng điểm trung vị hơn là bám sát biến động.

Đối với dự báo tổng thu ngân sách, khi mỗi khoản thu được dự báo theo mô hình tốt nhất, tổng thu ngân sách (tổng của các khoản thu thành phần) có: RMSE $\approx$ 36.33 nghìn tỷ, MAE $\approx$ 29.66 nghìn tỷ, $R^2 \approx 0.976$ so với tổng cấu phần thực. Sai số bình quân chỉ ở mức vài chục nghìn tỷ, tương đương cỡ 1-2% tổng thu trong những năm gần đây. Kết quả này cho thấy mô hình "lai" được xây dựng đủ tốt để dự báo thu NSNN của Việt Nam trong giai đoạn tới.

**4. Kết luận và hàm ý chính sách**

Kết quả nghiên cứu cho thấy không tồn tại "một mô hình tối ưu cho mọi sắc thuế", mà cần danh mục mô hình chuyên biệt. Có thể ứng dụng:

- RF/XGB cho các sắc thu nội địa quy mô lớn (PIT, VAT, EXT);

- ElasticNet cho sắc thu tuyến tính hoặc dữ liệu ít (CIT, EMT, AID);

- OLS cho thu dầu thô và khoản thu khác ổn định.

Cấu hình này tạo nền tảng cho hệ thống dự báo NSNN cập nhật liên tục, có khả năng cảnh báo sớm biến động thu NSNN và mô phỏng tác động chính sách. Tuy nhiên, một số khoản thu như AID_R có $R^2$ thấp do bản chất hành chính và rời rạc; cần đặc tả thêm biến thể chế hoặc tín hiệu lịch phát hành - giải ngân ODA. Nhóm EMT_R/VATM_R chịu ảnh hưởng mạnh từ lịch trình cắt giảm thuế quan và biến động logistics; nên tăng biến thương mại theo ngành và chỉ số chi phí vận tải. Trong giai đoạn tới, có thể thử nghiệm mô hình hai giai đoạn dựa trên phần dư (two-stage residual modeling), kết hợp mô hình tuyến tính và phi tuyến để tiếp tục giảm sai số.

Tổng thể, việc ứng dụng học máy trong dự báo thu NSNN đã chứng minh giá trị vượt trội: cải thiện độ chính xác, tăng tính thích ứng và tạo tiền đề cho quản trị tài khóa chủ động trong kỷ nguyên dữ liệu lớn và chuyển đổi số tài chính công.

**Tài liệu tham khảo:**

*Department of Treasury and Finance Victoria. (2024). Applying Machine Learning in Tax Revenue Forecasting. Melbourne, Australia. Retrieved from https://www.dtf.vic.gov.au/victorias-economic-bulletin-applying-machine-learning-tax-revenue-forecasting*

*World Bank. (2023). A Strategic AI Integration Model for Revenue Administrations. Washington, D.C. Retrieved from https://documents1.worldbank.org/curated/en/099071025155040812/pdf/P505930-22d00dc3-5bb0-4f32-8adf-76e8622fbd1d.pdf*

*Inland Revenue Authority of Singapore (2023). AI-driven Forecasting for Goods and Services Tax (GST) Collections. Singapore.*

*Ministry of Finance, India (2023). AI-based GST Revenue Forecasting using Long Short-Term Memory Networks. New Delhi. Retrieved from https://www.tandfonline.com/doi/full/10.1080/23322039.2023.2285649#abstract*