XÂY DỰNG MÔ HÌNH GỢI Ý GIÁ BÁN CHO NGƯỜI MỚI TRÊN CÁC SÀN TMĐT

BUILDING A PRICE SUGGESTION MODEL FOR NEWCOMERS ONE-COMMERCE PLATFORMS

***Hồ Văn Như Ý*[1]**

[1]Khoa Công Nghệ Thông Tin, Trường Đại học Kinh Tế Tài Chính, thành phố Hồ Chí Minh, Việt Nam, yhvn24@uef.edu.vn

***Tóm tắt**: Trong bối cảnh thị trường thương mại điện tử ngày càng cạnh tranh, việc tối ưu hóa giá sản phẩm trở thành yếu tố quan trọng. Nghiên cứu này đề xuất một mô hình gợi ý giá sản phẩm bằng các mô hình Neural Networks, XGBoost, Stacked Regressor, Grid Search và Simple Regressor. Kết quả cho thấy, mô hình học máy Neural Network với 3 hidden layers, Stacked Regressor và Simple Regressor đều cho kết quả dự báo và hiệu suất tốt. Các thông số đầu vào quan trọng để gợi ý giá sản phẩm trên sàn TMĐT thông qua các mô hình học máy bao gồm thương hiệu, chỉ số đánh giá, chỉ số số lượng sản phẩm đã bán, chỉ số giá niêm yết và chỉ số giá gốc.*

***Từ khóa**: Gợi ý giá sản phẩm, Nền tảng thương mại điện tử, mạng Neural, Mô hình học máy, XGBoost.*

***Abstract**: In the context of an increasingly competitive e-commerce market, optimizing product prices has become a crucial factor. This study proposes a product price suggestion model using Neural Networks, XGBoost, Stacked Regressor, Grid Search, and Simple Regressor models. The results show that the Neural Network machine learning model with 3 hidden layers, XGBoost, Stacked Regressor, Grid Search, and Simple Regressor all yield good forecasting results and performance. The important input parameters for suggesting product prices on e-commerce platforms through machine learning models include brand, rating index, number of products sold index, listed price index, and original price index.*

***Keywords**: Product price suggestion, E-commerce platform, Neural network, Machine learning model, XGBoost.*

1. Đặt vấn đề

Dựa trên bối cảnh phát triển mạnh mẽ của các sàn thương mại điện tử (TMĐT) như Shopee, Lazada, Tiki, Amazon, Alibaba,… cùng với sự bùng nổ của công nghệ như AI,

Blockchain và smartphone, việc mua bán trực tuyến ngày càng trở nên phổ biến và đóng vai trò quan trọng trong đời sống kinh tế xã hội. TMĐT không chỉ mang lại sự tiện lợi cho người tiêu dùng mà còn mở ra cơ hội kinh doanh lớn cho các nhà bán hàng, từ các cửa hàng nhỏ lẻ đến các doanh nghiệp lớn. Tuy nhiên, bên cạnh tiềm năng to lớn này, thực trạng cạnh tranh gay gắt giữa các cửa hàng, bao gồm cả cạnh tranh lành mạnh và không lành mạnh, đang tạo ra thách thức lớn trong việc đảm bảo sự ổn định giá sản phẩm trên thị trường. Hiện nay, các nhà bán hàng thường gặp khó khăn trong việc xác định mức giá hợp lý để vừa đảm bảo khả năng cạnh tranh, vừa duy trì lợi nhuận, đặc biệt trong bối cảnh giá cả biến động liên tục và hành vi mua sắm của người tiêu dùng thay đổi nhanh chóng. Sự thiếu minh bạch trong cơ chế định giá cùng với việc chạy đua giảm giá để thu hút khách hàng còn tiềm ẩn nguy cơ thua lỗ lớn, đặc biệt đối với các cửa hàng vừa và nhỏ. Điều này không chỉ làm suy giảm hiệu quả kinh doanh mà còn ảnh hưởng đến niềm tin và trải nghiệm mua sắm của người tiêu dùng trên các sàn TMĐT[1][2].

Vấn đề nghiên cứu gợi ý giá bán sản phẩm bằng các phương pháp học máy nhằm giải quyết thực trạng trên trở nên vô cùng cấp thiết. Việc sử dụng các thuật toán học máy như Linear Regression, Support Vector Regression[3], Decision Tree hay các mô hình học sâu (deep learning) hiện đại có thể giúp phân tích khối lượng lớn dữ liệu liên quan đến thị trường, hành vi mua sắm của người tiêu dùng, lịch sử giá sản phẩm và yếu tố cạnh tranh giữa các cửa hàng. Những mô hình này sẽ đưa ra các gợi ý giá tối ưu, phù hợp với biến động của thị trường cũng như đặc điểm riêng của từng sản phẩm và nhóm khách hàng mục tiêu. Không chỉ giúp người bán tiết kiệm thời gian và chi phí trong việc định giá sản phẩm, giải pháp này còn góp phần hạn chế tình trạng giảm giá quá mức, duy trì lợi nhuận ổn định và nâng cao năng lực cạnh tranh của các cửa hàng trên sàn TMĐT. Đồng thời, việc áp dụng công nghệ học máy để gợi ý giá còn giúp các sàn TMĐT xây dựng một môi trường kinh doanh minh bạch, lành mạnh và bền vững, từ đó tạo niềm tin cho người tiêu dùng và thúc đẩy sự phát triển ổn định của thị trường TMĐT trong tương lai. Nghiên cứu này sẽ tập trung vào việc xây dựng mô hình dự đoán và gợi ý giá bán hiệu quả, có khả năng thích ứng với những thay đổi của thị trường và nhu cầu của các bên liên quan, mang lại lợi ích thiết thực cho cả người bán, người tiêu dùng và các sàn TMĐT.

2. Tổng quan nghiên cứu

Giá là giá trị số tiền đó được xem là chi phí ước tính của một cái gì hoặc một vật cụ thể nào đó trên thị trường, cho dù đó là một sản phẩm, hay nó là một dịch vụ cụ thể[4]. Giá cả là một

khái niệm dùng để chỉ số tiền hoặc giá trị khác mà người mua hoặc sử dụng sản phẩm/ dịch vụ phải trả cho người bán. Giá cả được xác định bởi một số yếu tố, bao gồm chi phí sản xuất, cung cầu, chi phí vận chuyển, thuế và lợi nhuận, các sản phẩm cạnh tranh, mức cung cầu,… Đặc biệt, trong thời kì phát triển mạnh của công nghệ thông tin, của sàn TMĐT, nên đã thu hút được rất nhiều nhà nghiên cứu quan tâm. Sự thay đổi của giá sản phẩm không chỉ trên thị trường "truyền thống" mà còn thay đổi trên các sàn TMĐT có liên quan đến sự phát triển của nền kinh tế trong nước theo cả chiều sâu và chiều rộng, lẫn tích cực và tiêu cực (BNEWS.VN). Các yếu tố này gây ra sự biến động của giá sản phẩm và sự khó khăn trong đánh giá xu hướng biến động của nó. Một số phương pháp dự báo truyền thống có thể kể đến như: LR[5], RF và DT. Theo đó, giá sản phẩm gợi ý được xác định bởi các mức giá bán của sản phẩm, số lượng sản phẩm đã bán, mức đánh giá sản phẩm, thương hiệu của sản phẩm,…Tuy nhiên, với sự phát triển của trí tuệ nhân tạo (AI), các phương pháp học máy như: SVM, mạng nhân tạo (ANN), DT, XGBoost[5], RF, mạng bộ nhớ ngắn, dài hạn đã được sử dụng để dự báo sự thay đổi giá sản phẩm.

Các mô hình truyền thống như ARIMA thường được đánh giá cao về tính đơn giản và khả năng giải thích, nhưng hạn chế trong việc xử lý dữ liệu lớn và phi tuyến tính. Trong khi đó, học máy như XGBoost hay ANN cung cấp độ chính xác cao nhưng cần khả năng xử lý tính toán lớn.

ErenElagz đã phát triển một mô hình dự báo giá sản phẩm dựa trên mô hình LSTM Neural Networks và nhận thấy rằng mô hình trên là phương pháp tốt nhất cho số lượng lớn tập dữ liệu và lựa chọn thuộc tính tương ứng.

Nghiên cứu của Nguyễn Thái Sơn (2023) bằng các phương pháp ANFIS, ANN, GMDH, LSTM đã chỉ ra rằng, biến động giá của các mặt hàng, đặc biệt với các mặt hàng chủ chốt ảnh hưởng trực tiếp đến nhiều lĩnh vực của nền kinh tế, chính sách đầu tư và vận hành các sàn TMĐT – dự đoán giá, vận hành dịch vụ và chuỗi cung ứng[6],…

Theo báo cáo thị trường sàn TMĐT của Metric.vn[7], giá sản phẩm là yếu tố then chốt ảnh hưởng đến doanh thu và sức cạnh tranh của nhà bán trên sàn TMĐT, vì việc lựa chọn phân khúc giá phù hợp và áp dụng chiến lược giá linh hoạt giúp tối ưu hóa hiệu quả kinh doanh của cửa hàng[8]. Tuy nhiên, sàn thương mại điện tử sẽ khó quản lí, ổn định thị trường khi có sự cạnh tranh khốc liệt về giá bán sản phẩm. Đồng thời, khách hàng phải dành nhiều thời gian để tìm các mặt hàng tương tự và so sánh về giá sản phẩm. Không chỉ khách hàng và sàn TMĐT, biến động giá sản phẩm cũng ảnh hưởng đến lợi nhuận, sức tồn tại của các cửa hàng, nhà bán nhỏ, lẻ, dẫn đến sự rút lui khỏi thị trường.

Roland Szilágyi, Beatrix Varga, và Renata Geczi-Papp (2016) so sánh hiệu suất dự báo của các phương pháp trung bình trượt (Moving Average), xu hướng phân tích (Analytical Trend), làm mịn hàm mũ (Exponential Smoothing), Box-Jenkins (ARIMA) và xu hướng trượt với trọng số điều hòa, kết quả chỉ ra hiệu suất dự báo vượt trội hơn với độ chính xác cao và không có sai lệch hệ thống đáng kể của xu hướng trượt với trọng số điều hòa so với các phương pháp học máy khác[9].

3. Cơ sở lý thuyết

Nghiên cứu sử dụng các mô hình: Simple Regressor (SR), Grid Search (GS), Stacked Regressor, XGBoost và Neural Network để gợi ý giá sản phẩm dựa trên dữ liệu của Shopee được thu thập từ APIs. Vì có thể nói, Shopee là một trong những sàn TMĐT phổ biến nhất tại Đông Nam Á và Việt Nam, với lượng giao dịch lớn và cơ sở dữ liệu phong phú. Ngoài ra, mô hình Stacked Regressor kết hợp nhiều mô hình như XGBoost và Neural Network để đạt độ chính xác cao nhất trong việc đề xuất giá bán cho sản phẩm thể loại có doanh thu cao trên Shopee.

3.1. Simple Regressor (SR)

Mô hình SR[10] là một công cụ quan trọng trong thống kê và học máy, được sử dụng để đánh giá các giá trị thực dựa trên các biến đầu vào và mô hình hóa mối quan hệ giữa biến độc lập và biến phụ thuộc bằng cách đưa ra một giải pháp thích hợp. Với biến phụ thuộc là giá cuối cùng, mô hình SR thiết lập mối quan hệ giữa giá cuối cùng, các yếu tố tác động bởi phương trình hồi quy và tối thiểu hóa tổng bình phương sai số (RSS):

- final_price $= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$
- Trong đó:

  final_price là biến phụ thuộc (dependent variable), biến giá bán gợi ý cho người bán.

  X_1, X_2, X_n, n là biến độc lập, hay còn gọi là biến giải thích được thể hiện trong Bảng 1, các biến này là các yếu tố đầu vào có ảnh hưởng đến giá trị của final_price.

  β_0: Hệ số chặn (intercept), đại diện cho giá trị của final_price khi tất cả các biến độc lập đều bằng 0.

  $\beta_1, \beta_2, ..., \beta_n$: Các hệ số hồi quy (regression coefficients), biểu diễn mức độ ảnh hưởng của mỗi biến độc lập $X_1, X_2, ..., X_n$ lên final_price.

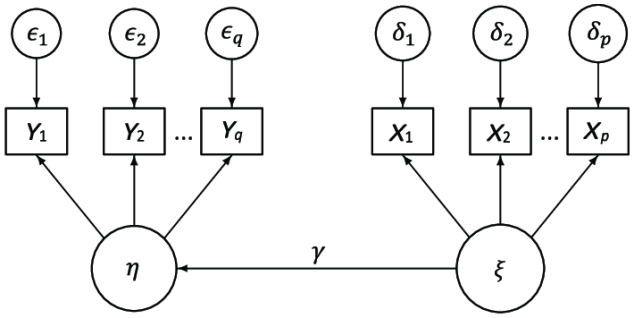
  ε: Sai số ngẫu nhiên (error term), đại diện cho những yếu tố không được giải thích bởi các biến độc lập trong mô hình.

- RSS=$\sum_{i=1}^{m}$*(final_price$^{(i)}$−final_price$^{\wedge(i)}$)^2
  - RSS đo lường tổng bình phương phần dư (chênh lệch giữa giá trị thực tế và giá trị dự đoán), dùng để đánh giá mức độ phù hợp của mô hình với dữ liệu đầu v
  - Trong đó:

    m là số lượng mẫu dữ liệu đầu vào.

    final_price$^{(i)}$ là giá trị thực tế của biến phụ thuộc tại mẫu thứ i

    final_price$^{\wedge(i)}$ là giá trị dự đoán của mô hình tại mẫu thứ i

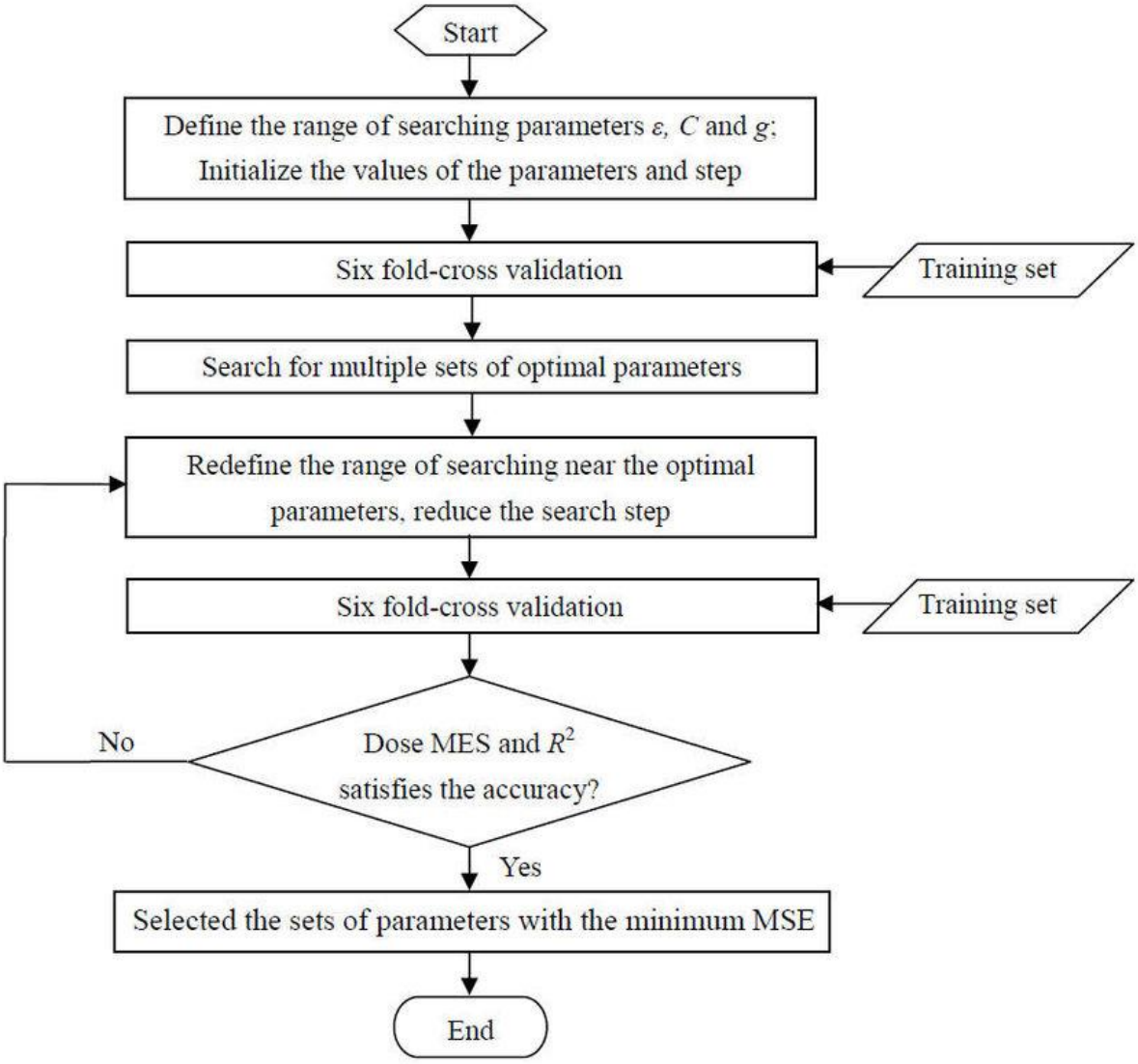


Hình 1: Model Simple Regressor

3.2. Grid Search

Grid Search[11] (GS) là một kỹ thuật tìm kiếm siêu tham số rất phổ biến trong học máy do khả năng sử dụng và hiệu quả của mô hình. Mô hình thực hiện tìm kiếm hoàn chỉnh trên một tập hợp con nhất định của không gian siêu tham số của thuật toán đào tạo. Tìm kiếm lưới sẽ

gặp khó khăn với không gian có nhiều chiều, nhưng thường có thể dễ dàng song song hóa, vì các giá trị siêu tham số mà thuật toán làm việc thường đối lập với nhau.



Hình 2: Grid Search Model

- Trong đó:

  ε (epsilon): Tham số liên quan đến độ chính xác hoặc độ chấp nhận sai số.

  C: Tham số điều chỉnh mức độ phạt (regularization parameter), thường được sử dụng trong các thuật toán như SVM để điều chỉnh mức độ phạt cho các lỗi phân loại.

  g (gamma): Tham số trong hàm kernel (trong SVM hoặc các thuật toán khác), điều khiển độ ảnh hưởng của các điểm dữ liệu.

  Training set: Bộ dữ liệu được chia theo tỉ lệ dùng để huấn luyện mô hình trong từng bước kiểm tra và tối ưu.

  MSE (Mean Squared Error): Trung bình phương sai số, dùng để đánh giá sai số giữa giá trị dự đoán và giá trị thực tế.

- MSE = $\frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$

- Trong đó:

  N là số các quan sát (điểm dữ liệu)

  $\frac{1}{N}\sum_{i=1}^{n}$ là giá trị trung bình

  y_i là giá trị thực quan sát được và $mx_i + b$ (final_price) là giá trị dự đoán.

- R^2 (R – squared): Hệ số xác định, do mức độ phù hợp của mô hình với dữ liệu.

  - $R^2 = 1 - \frac{ESS}{TSS}$

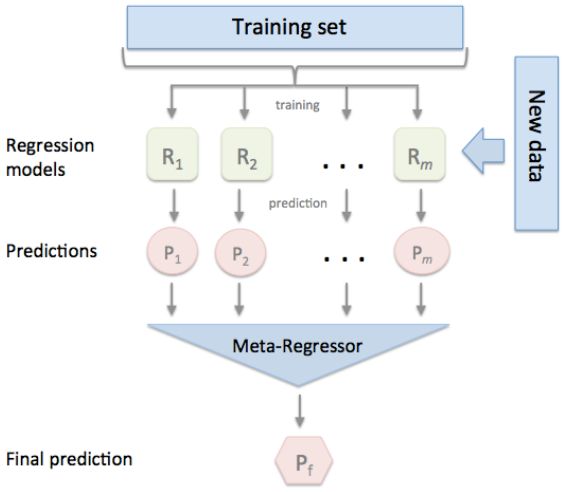
- Trong đó:

  ESS (Residual Sum of Squares) là tổng các độ lệch bình phương của phần dư.

  TSS (Total Sum of Squares) là tổng độ lệch bình phương của toàn bộ các nhân tố nghiên cứu.

3.3. Stacking Regressor

Mô hình Stacked Regressor[12][13] là một kỹ thuật meta – learning trong học máy, trong đó, ta kết hợp nhiều mô hình cơ bản (base leaners) để tạo ra một mô hình hiệu quả và mạnh mẽ hơn. Mô hình này sử dụng một "meta – model" để học cách kết hợp đầu ra từ các mô hình cơ bản, khai thác điểm mạnh của từng mô hình.



Hình 3: Stacking Regressor model

- Trong đó:

  Các biến đầu vào chính là các đặc trưng (features) của tập dữ liệu huấn luyện ban đầu (Training Set).

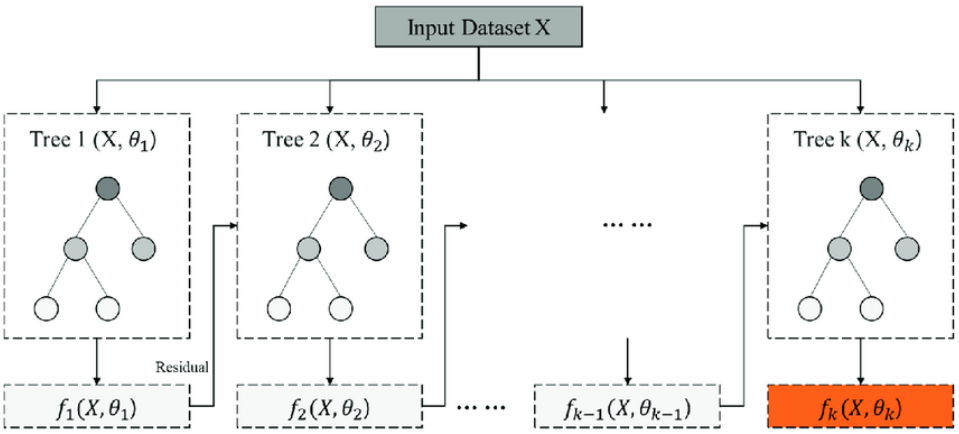
  Các mô hình hồi quy cơ sở $R_1, R_2, \ldots, R_m$ được huấn luyện độc lập để dự đoán giá trị mục tiêu.

P_m là những giá trị được tạo ra từ các mô hình hồi quy cơ sở R_m và được sử dụng làm các biến đầu vào cho mô hình meta (Meta – Regressor).

P_f là giá trị mục tiêu cuối cùng.

3.4. XGBoost

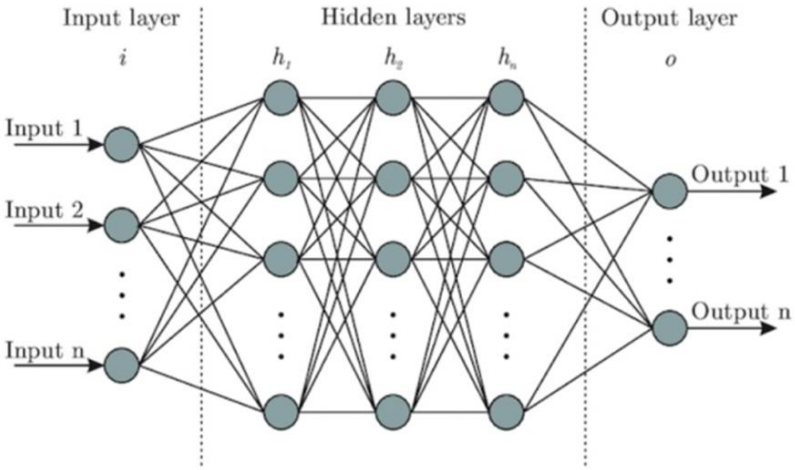
Mô hình XGBoost[14] là một kỹ thuật học máy thuộc danh mục học tập tổng hợp, cụ thể là khung tăng cường độ dốc. Nó sử dụng cây quyết định làm người học cơ sở và sử dụng các kỹ thuật chính quy hóa để nâng cao khả năng khái quát hóa mô hình. XGBoost có các tính năng hoạt động như: chính quy hóa, xử lí các dữ liệu thưa thớt, phác thảo lượng tử có trọng số, cấu trúc khối để học song song, nhận biết bộ đệm, điện toán ngoài lõi. Các thông số chung là: booster, silent, nthread. Ngoài ra, còn các thông số tăng cường giúp mô hình đạt kết quả cao hơn như: eta (tốc độ học của mô hình), min_child_weight (số lượng tối thiểu của lá có trong cây), max depth (độ sâu tối đa của mỗi cây quyết định trong ensemble), gamma, lambda,…



Hình 4: XGBoost model process

3.5. Neural Network

Mô hình Neural Network[15] (NN) là một cấu trúc tính toán lấy cảm hứng từ cách hoạt động của bộ não con người, được sử dụng để học từ dữ liệu và thực hiện các tác vụ như phân loại, hồi quy, và nhận diện mẫu.



Hình 5: Cấu trúc Neural Network

3.6. Bộ dữ liệu

Nghiên cứu dựa trên bộ dữ liệu shopee-dataset-samples của Bright Data được thu thập từ APIs của sàn thương mại điện tử Shopee. Bộ dữ liệu gồm 898 dòng và 17 trường dữ liệu. Các biến sử dụng trong mô hình được miêu tả ở Bảng 1.

Bảng 1:Mô tả các biến trong nghiên cứu

Biến	Mô tả	Nguồn
product_id	Mã định danh duy nhất cho sản phẩm trên Shopee.	APIs
product_name	Tên sản phẩm.	APIs
variation	Thông tin về các biến thể sản phẩm.	APIs
category_id	Mã định danh phân loại danh mục sản phẩm.	APIs
brand	Thương hiệu của sản phẩm, ảnh hưởng mạnh đến quyết định mua hàng do sự nhận diện thương hiệu và lòng tin từ khách hàng.	APIs
rating	Đánh giá trung bình của khách hàng cho sản phẩm, chỉ số quan trọng phản ánh chất lượng sản phẩm.	APIs
initial_price	Giá gốc hoặc giá khởi điểm của sản phẩm.	APIs
final_price	Giá hiện tại hoặc giá cuối cùng của sản phẩm sau khi giảm giá hoặc khuyến mại (là giá trị mục tiêu của mô hình).	APIs
currency	Loại tiền tệ.	APIs
sold	Số lượng sản phẩm đã bán, biểu thị mức độ phổ biến của sản phẩm và đóng vai trò quan trọng trong việc dự đoán giá.	APIs

status	Tình trạng sản phẩm.	APIs
image	Hình ảnh sản phẩm.	APIs
flash_sale	Giảm giá sản phẩm.	APIs
vouchers	Phiếu giảm giá.	APIs
category	Tết mục phân loại sản phẩm.	APIs
price_to_sold_ratio	Tỉ lệ giá lượng (giá/số lượng) là chỉ số kết hợp giữa giá và số lượng bán, giúp đánh giá hiệu quả giá trị của sản phẩm trên thị trường.	Jupyter

4. Kết quả và thảo luận

Bảng 2 biểu diễn các biến độc lập và phụ thuộc trong giai đoạn nghiên cứu, cho thấy mức giảm giá dao động lớn, với tỷ lệ giảm giá lớn nhất lên đến 90% và chênh lệch giá trị tuyệt đối trung bình là $5.70. Biến initial_price và final_price thể hiện sự khác biệt đáng kể về mức giá trước và sau giảm, làm nổi bật tác động rõ rệt của chính sách giá đến doanh số bán hàng **(**sold**)** và lợi nhuận cuối cùng.

Ngoài các yếu tố định lượng trong mô hình, các yếu tố phi giá như dịch vụ chăm sóc khách hàng, chính sách giao hàng nhanh, hoặc trải nghiệm mua sắm cũng đóng vai trò quan trọng. Ví dụ, việc cung cấp dịch vụ giao hàng trong ngày hoặc chính sách đổi trả linh hoạt có thể ảnh hưởng đến giá trị cảm nhận của người dùng, và từ đó gián tiếp ảnh hưởng đến quyết định mua hàng và giá bán.

Phân tán lớn ở cả giá gốc **(**initial_price**)** với std = 169.89 và giá sau giảm **(**final_price**)** với std = 164.63 cho thấy sự đa dạng trong chiến lược định giá giữa các danh mục sản phẩm, có thể phản ánh các yếu tố như chất lượng, thương hiệu, hoặc mức độ đáp ứng nhu cầu thị trường. Đồng thời, giá sản phẩm sau giảm tiếp tục cho thấy mức độ phân tán cao, đặc biệt ở các sản phẩm có doanh số lớn.

Tỷ lệ giá trên số lượng bán **(**price_to_sold_ratio**)** rất thấp ở phần lớn sản phẩm, chỉ ra rằng các sản phẩm có doanh số cao thường được bán ở mức giá thấp. Điều này cho thấy chiến lược giảm giá mạnh có thể đóng vai trò quan trọng trong việc thúc đẩy doanh số. Tuy nhiên, chiến lược này cũng làm giảm giá trị trung bình của sản phẩm, đặt ra thách thức về việc cân đối giữa tăng trưởng doanh số và duy trì lợi nhuận tổng thể.

Dữ liệu sau khi được chia ngẫu nghiên thành hai tập huấn luyện và tập kiểm tra theo tỉ lệ 70% và 30% sau khi đã làm sạch với mục đích để huấn luyện mô hình và kiểm tra để đánh giá hiệu quả của mô hình.

Bảng 2: Thống kê mô tả

Biến	**product_id**	**category_id**	**initial_price**	**final_price**	**sold**	**price_to_sold_ratio**
count	8.98E+02	898	898	898	8.98E+02	898
mean	2.02E+10	100413.5668	44.938419	39.241091	1.05E+05	3.908293
std	5.90E+09	293.293044	169.890234	164.632516	1.22E+06	66.2979
min	1.00E+08	100001	0	0	0.00E+00	0
25%	1.71E+10	100015	9.05	6.7	1.00E+02	0.000623
50%	2.21E+10	100630	19.1	16.15	2.14E+03	0.006169
75%	2.46E+10	100637	36.5	32.225	2.08E+04	0.12899
max	2.99E+10	100644	3935	3935	3.45E+07	1967.5

Dữ liệu về giá sản phẩm và các biến giải thích sau khi được chia ngẫu nhiên thành hai tập huấn luyện và kiểm tra được sử dụng trong hai giai đoạn: giai đoạn thứ nhất huấn luyện mô hình từ tập dữ liệu huấn luyện với sáu phương pháp là Simple Regressor, Grid Search, Stacked Regressor, Neural Network với 3 hidden layers, Neural Network với 5 hidden layers, Neural Network với 10 hidden layers và XGBoost; giai đoạn thứ hai gợi ý sử dụng mô hình đã được huấn luyện trên tập dữ liệu kiểm tra. Các giá trị đánh giá mô hình trên tập kiểm tra thể hiện trong Bảng 3.

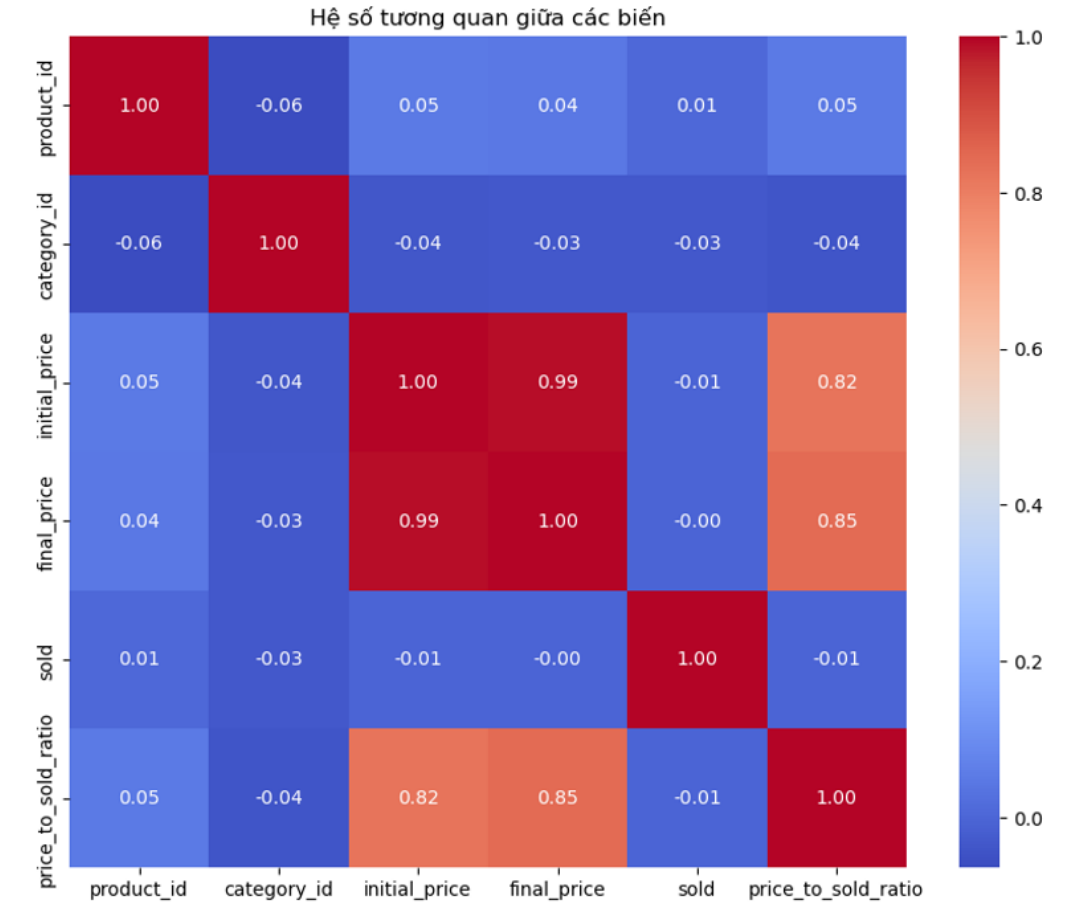
Bảng 3:Giá trị đánh giá

Mô hình	MAE		MSE	RMSE	Learning_rate	Max_depth	N_estimators	Best core
	train	test						
Simple Regressor	0.038	2.565	x	x	x	x	x	x
Grid Search	x	x	7.92	2.81	0.07	7	500	-7.9249
Stacked Regressor	x	2.223	x	x	x	x	x	x
Neural Network với 3	1.948	1.602	x	x	x	x	x	x

hidden layers								
Neural Network với 5 hidden layers	3.909	3.913	x	x	x	x	x	x
Neural Network với 10 hidden layers	15.826	18.462	x	x	x	x	x	x

Bảng 3 cho thấy, ba mô hình Neural Network (3 hidden layers), Stacked Regressor và Grid Search đều cho kết quả dự báo tốt hơn so với ba mô hình còn lại. Mô hình Neural Network (3 hidden layers) cho kết quả dự báo chính xác nhất với test_mae là 1.602, tốt nhất trong các mô hình mà không gặp hiện tượng overfiting với cấu hình vừa phải (với 3 hidden layers có cấu trúc tối giản). Với Stacked Regressor và Grid Search cung cấp hiệu suất ổn định, phù hợp khi cần tinh chỉnh kỹ thuật. Có thể một số biến độc lập trong mô hình có tương quan cao với nhau làm ảnh hưởng đến hiệu của của mô hình Neural Network với 10 hidden layers, Simple Regressor, Neural Network với 5 hidden layers. Do đó mô hình Neural Network (3 hidden layers) có thể được sử dụng dể gợi ý giá sản phẩm trong thời gian kế tiếp.

Từ kết quả gợi ý đánh giá mô hình Neural Network (3 hidden layers), nghiên cứu thực hiện đánh giá tầm quan trọng của các biến trong mô hình. Các biến, initial_price, final_price, sold và price_to_sold đóng vai trò quan trọng nhất trong gợi ý giá sản phẩm, hay nói cách khác, các biến này có tác động đến giá sản phẩm trên sàn Shopee. Tiếp theo là các biến đại diện cho chỉ số đánh giá (“rating”) cũng tác động đến giá sản phẩm nhưng với mức biến động không đáng kể.



Biểu đồ 1: Hệ số tương quan giữa các biến

Dựa trên biểu đồ hệ số tương quan giữa các biến, ta nhận thấy rằng biến initial_price và final_price có hệ số tương quan cực kỳ cao (0.99), cho thấy mối quan hệ gần như tuyến tính hoàn hảo. Điều này gợi ý rằng hai biến này mang thông tin tương tự như nhau, và có thể cân nhắc loại bỏ một trong hai để giảm sự dư thừa của dữ liệu, đồng thời đơn giản hóa mô hình và giảm nguy cơ đa cộng tuyến. Bên cạnh đó, biến price_to_sold_ratio có tương quan mạnh với cả initial_price (0.82) và final_price (0.85), điều này có thể gây ra hiện tượng đa cộng tuyến khi sử dụng đồng thời trong mô hình với các biến giá. Do đó, cần cân nhắc kỹ lưỡng khi đưa biến này vào phân tích. Ngoài ra, các biến như initial_price và final_price có tương quan rất thấp với sold, cho thấy mối quan hệ không rõ ràng hoặc không tuyến tính giữa các biến này và số lượng hàng bán. Điều này gợi ý rằng cần thử nghiệm thêm các phương pháp phân tích. Ngoài ra, các biến như initial_price và final_price có sự tương quan rất thấp với sold, cho thấy mối quan hệ không rõ ràng hoặc không tuyến tính giữa các biến này và số lượng hàng bán. Điều này gợi ý rằng cần thử nghiệm thêm các phương pháp phân tích phi tuyến tính để đánh giá vai trò của giá đến kết quả bán hàng. Cuối cùng, các biến như product_id và category_id không có mối tương quan đáng kể với các biến khác, phù hợp với vai trò nhận dạng và phân loại. Từ những

phân tích trên, có thể xem xét loại bỏ các biến dư thừa hoặc chuyển sang các phương pháp phân tích phù hợp hơn để cải thiện hiệu quả của mô hình.

5. Kết luận và hướng phát triển

Các mô hình học máy đã được chứng minh là hiệu quả hơn so với các mô hình thống kê truyền thống trong nhiều lĩnh vực khác nhau, nhờ vào ưu điểm vượt trội không yêu cầu chặt chẽ về điều kiện dữ liệu và mối quan hệ giữa các biến độc lập với biến phụ thuộc. Nhờ vậy, các mô hình học máy có khả năng ứng dụng rộng hơn và linh hoạt hơn. Trong nghiên cứu này, mô hình Neural Network với ba lớp ẩn (3 hidden layers) đã cho kết quả vượt trội so với các mô hình Stacked Regressor và Grid Search trong việc gợi ý giá sản phẩm trên nền tảng Shopee. Đồng thời, nghiên cứu đã xác định các yếu tố tác động đến giá sản phẩm thông qua việc đánh giá tầm quan trọng của các đặc tính trong mô hình hồi quy. Điều này mang lại giá trị thực tiễn lớn cho các cửa hàng và sàn TMĐT, giúp họ đánh giá chính xác sự biến động giá sản phẩm và xây dựng các chiến lược kinh doanh hiệu quả hơn.

Tuy nhiên, để nâng cao hiệu quả dự báo của các mô hình học máy, cần liên tục bổ sung và cập nhật thông tin, dữ liệu mới nhằm bắt kịp những thay đổi hoặc chuyển biến nhanh chóng của thị trường. Trong một số trường hợp, việc xây dựng lại mô hình dự báo có thể là cần thiết để đảm bảo phù hợp với tình hình mới. Ngoài ra, cần cập nhật và phân tích các yếu tố liên quan như thông tin về thuế, chi phí vận chuyển, và các chính sách kinh doanh. Mặc dù những yếu tố này khó có thể tích hợp trực tiếp vào mô hình học máy, nhưng chúng đóng vai trò quan trọng trong việc định hướng và tránh sai lệch trong việc đưa ra quyết định dựa trên kết quả dự báo.

Nghiên cứu này đã mang lại những đóng góp quan trọng cả về mặt khoa học lẫn thực tiễn trong lĩnh vực TMĐT:

Về mặt khoa học, nghiên cứu đã giới thiệu một cách tiếp cận toàn diện để phân tích mối quan hệ giữa các yếu tố định giá sản phẩm, tỷ lệ bán hàng và các đặc điểm phân loại khác trong ngành TMĐT. Việc xây dựng và đánh giá mô hình không chỉ giúp phát hiện các mối quan hệ quan trọng giữa các biến, mà còn chỉ ra những yếu tố dư thừa, từ đó tối ưu hóa quy trình phân tích dữ liệu và nâng cao hiệu quả mô hình hóa. Đồng thời, nghiên cứu cũng mở ra tiềm năng ứng dụng trong các lĩnh vực khác có cấu trúc dữ liệu tương tự.

Về mặt thực tiễn, kết quả nghiên cứu mang lại nhiều ứng dụng thiết thực trong TMĐT. Thứ nhất, mô hình có thể hỗ trợ tối ưu hóa chiến lược định giá sản phẩm, giúp doanh nghiệp xác định mức giá tối ưu để tối đa hóa doanh thu hoặc số lượng bán ra. Thứ hai, các phát hiện về mối quan hệ giữa giá và tỷ lệ bán hàng cho phép doanh nghiệp dự báo chính xác hiệu quả của các chương trình khuyến mãi hoặc các chiến lược điều chỉnh giá. Thứ ba, việc nhận diện các yếu tố quan trọng trong mô hình giúp doanh nghiệp cá nhân hóa trải nghiệm khách hàng, nâng cao tỷ lệ chuyển đổi và mức độ hài lòng của người mua.

Ngoài ra, nghiên cứu cũng đưa ra các giải pháp giảm thiểu hiện tượng đa cộng tuyến trong phân tích dữ liệu, từ đó cải thiện độ chính xác và tính ổn định của mô hình trong các bài toán dự báo. Với sự phát triển nhanh chóng của TMĐT, việc xây dựng hệ thống tự động gợi ý giá bằng trang web hoặc ứng dụng từ dữ liệu thị trường, dữ liệu chương trình khuyến mãi, trải nghiệm khách hàng,…Những ứng dụng từ nghiên cứu này có tiềm năng giúp các doanh nghiệp không chỉ cải thiện năng lực cạnh tranh mà còn thích ứng nhanh với những thay đổi trên thị trường. Trong tương lai, việc kết hợp học máy với AI sinh tạo (Generative AI) để tự động phân tích thị trường sẽ giúp tối đa hoá việc đề xuất giá trong thời gian thực trên các sàn TMĐT.

Tài liệu tham khảo

[1] Metric – Nền tảng số liệu E - commerce, *Báo cáo Thị trường TMĐT quý I/2023 và dự báo quý II/2023.*

[2] *Áp dụng thương mại điện tử: Nghiên cứu trường hợp các doanh nghiệp nhỏ và vừa Việt Nam (2023).*

[3] *Nguyễn Đức Hiền, Mô hình dự đoán giá cổ phiếu với K – Means và Fuzzy – SVM (2014).*

[4] TS. Hồ Quế Hậu (2023), *Bản chất và các yếu tố cấu thành giá trị hàng hóa,* Trường Đại học Kinh tế TP.Hồ Chí Minh.

[5] Phan Anh, Trương Thị Thùy Dương, Dự báo giá vàng nhìn từ mô hình hồi quy tuyến tính và mô hình học máy.

[6] Nguyễn Thái Sơn, Nghiên cứu ứng dụng kỹ thuật trí tuệ nhân tạo trong bài toán dự báo giá của một số mặt hàng (2023).

[7] Metric – Nền tảng số liệu E - commerce, *Báo cáo Thị trường TMĐT quý I/2023 và dự báo quý II/2023.*

[8] *Dr.Manjula Jain, A review on the variables that influence product pricing and decision (2021).*

[9] *Szilágyi Roland1, Varga Beatrix2, Géczi-Papp Renáta3, Possible methods for price forecating (2016).*

[10] *Tianyuan Guan, Mohammed Khorshed Alam, Marepalli Bhaskara Rao, Sample Size Calculations in Simple Linear Regression (2023).*

[11] *Petro Liashchynskyi, Pavlo Liashchynskyi,Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS.*

[12] *David H.Wolpert, Stacked Density Estimation (1997).*

[13] *Joseph T. Ornstein, Staked Regression and Poststratification (2019).*

[14] *Chen, T. và C. Guestrin. Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.*

[15] *TS. Nguyễn Chính Kiên, Ứng dụng mạng nơron nhân tạo vào bài toán dự báo (2017).*