HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Phùng Thị Thu

ÁP DỤNG MÔ HÌNH LONG SHORT - TERM MEMORY TRONG DỰ ĐOÁN GIÁ NHÀ TẠI VIỆT NAM

Chuyên ngành: Khoa học máy tính
Mã số: 8.48.01.01

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI - NĂM 2026

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. Đỗ Thanh Hà

Phản biện 1: PGS.TS Đỗ Trung Tuấn

Phản biện 2: PGS.TS Nguyễn Hữu Quỳnh

Đề án tốt nghiệp được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 9 giờ 45 phút ngày 24 tháng 01 năm 2026

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

1. Lý do chọn đề tài

Trong những năm gần đây, thị trường bất động sản tại Việt Nam, đặc biệt là tại các đô thị lớn như Hà Nội, thành phố Hồ Chí Minh đã chứng kiến sự biến động mạnh mẽ cả về giá cả lẫn nhu cầu giao dịch. Việc dự đoán chính xác giá nhà không chỉ hỗ trợ các cá nhân và tổ chức trong việc ra quyết định mua bán, đầu tư mà còn có ý nghĩa quan trọng trong quy hoạch và hoạch định chính sách phát triển đô thị.

Tuy nhiên, giá nhà thường biến động phức tạp do chịu ảnh hưởng từ nhiều yếu tố như vị trí địa lý, thời điểm giao dịch, chính sách kinh tế và xu hướng thị trường, v.v. Do đó, việc xây dựng một mô hình dự đoán chính xác là một thách thức lớn đối với các nhà nghiên cứu và nhà phân tích dữ liệu.

Trong bối cảnh đó, các phương pháp truyền thống như hồi quy tuyến tính hay mô hình ARIMA bộc lộ hạn chế khi xử lý dữ liệu phi tuyến và phụ thuộc thời gian dài. Trong khi đó, các mô hình học sâu, đặc biệt là mạng Long Short-Term Memory (LSTM) đã chứng minh hiệu quả vượt trội trong việc xử lý chuỗi thời gian có tính dài hạn và nhiều yếu tố ảnh hưởng. Hiện nay, việc ứng dụng LSTM vào dự đoán giá nhà tại Việt Nam vẫn còn khá mới mẻ và chưa có nhiều nghiên cứu chuyên sâu.

Xuất phát từ nhu cầu thực tiễn cũng như khoảng trống trong nghiên cứu hiện nay, đề tài ***"Áp dụng mô hình Long Short-Term Memory trong dự đoán giá nhà tại Việt Nam"*** được lựa chọn nhằm mục tiêu khai thác tiềm năng của mô hình LSTM trong việc dự báo chuỗi thời gian phức tạp, đồng thời đề xuất một hướng tiếp cận mới mang tính ứng dụng cao. Kết quả nghiên cứu kỳ vọng sẽ đóng góp vào việc nâng cao độ chính xác trong dự đoán giá nhà, đồng thời mở rộng khả năng ứng dụng của học sâu trong lĩnh vực bất động sản tại Việt Nam trong thời đại số hóa.

2. Tổng quan về vấn đề nghiên cứu

Trong những năm gần đây, thị trường bất động sản Việt Nam đã chứng kiến sự tăng trưởng mạnh mẽ, đặc biệt tại các thành phố lớn. Giá nhà liên tục biến động do ảnh hưởng của nhiều yếu tố, bài toán dự đoán giá nhà vẫn tồn tại nhiều thách thức. Dữ liệu bất động sản thường mang tính chuỗi thời gian, có độ trễ và phụ thuộc vào các yếu tố kinh tế xã hội phức tạp. Các phương pháp truyền thống như hồi quy tuyến tính, mô hình ARIMA hay cây quyết định thường không đủ khả năng xử lý mối quan hệ phi tuyến và phụ thuộc dài hạn trong dữ liệu.

Trong bối cảnh đó, các mô hình học sâu (Deep Learning) đã mở ra hướng tiếp cận mới. Đặc biệt, mô hình LSTM - một biến thể của mạng nơ-ron hồi quy (RNN) được thiết kế để ghi nhớ thông tin dài hạn trong chuỗi dữ liệu, giúp cải thiện độ chính xác trong dự đoán. LSTM đã được ứng dụng thành công trong nhiều lĩnh vực như tài chính, y tế, dịch vụ và đang dần được quan tâm trong lĩnh vực bất động sản.

Tại Việt Nam, việc ứng dụng LSTM vào dự đoán giá nhà vẫn còn khá mới mẻ, chưa có nhiều nghiên cứu chuyên sâu. Do đó, đề tài này hướng đến việc xây dựng mô hình LSTM phù hợp với dữ liệu thực tế tại Việt Nam, từ đó nâng cao hiệu quả và tiềm năng ứng dụng trong thực tiễn.

3. Mục đích nghiên cứu

Mục tiêu tổng quát: Xây dựng và đánh giá hiệu quả của mô hình LSTM trong việc dự đoán giá nhà tại Việt Nam dựa trên dữ liệu chuỗi thời gian nhằm cải thiện độ chính xác so với các phương pháp truyền thống.

Mục tiêu cụ thể: (i) Phân tích đặc điểm dữ liệu giá nhà tại Việt Nam và các yếu tố ảnh hưởng đến biến động giá; (ii) Tiền xử lý dữ liệu và xây dựng tập dữ liệu phù hợp cho mô hình học sâu; (iii) Thiết kế và huấn luyện mô hình LSTM để dự đoán giá nhà theo chuỗi thời gian; (iv) So sánh hiệu suất của mô hình LSTM với các mô hình truyền thống như hồi quy tuyến tính, ARIMA; (v) Đề xuất hướng ứng dụng

mô hình vào thực tiễn như hỗ trợ ra quyết định đầu tư, định giá tài sản hoặc hoạch định chính sách.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: (i) Dữ liệu giá nhà theo chuỗi thời gian tại Việt Nam (bao gồm các thông tin như giá bán, diện tích, vị trí, thời điểm giao dịch và các yếu tố liên quan đến thị trường bất động sản); (ii) Mô hình học sâu LSTM (có khả năng ghi nhớ thông tin dài hạn, giúp dự đoán xu hướng giá nhà dựa trên các yếu tố lịch sử); (iii) Các yếu tố ảnh hưởng đến giá nhà (vị trí địa lý, biến động thị trường, thu nhập bình quân, tốc độ đô thị hóa, v.v.).

Phạm vi nghiên cứu: (i) Không gian: Tập trung vào dữ liệu giá nhà đất tại thành phố Hà Nội, nơi có thị trường sôi động và dữ liệu tương đối đầy đủ; (ii) Thời gian: Dữ liệu được thu thập trong vòng khoảng 12 tháng nhằm đảm bảo tính cập nhật và phản ánh xu hướng thị trường; (iii) Kỹ thuật: Nghiên cứu tập trung vào việc xây dựng, huấn luyện và đánh giá mô hình LSTM. Không đi sâu vào các mô hình hybrid phức tạp hay các yếu tố pháp lý - chính sách của thị trường bất động sản; (iv) Dữ liệu: Sử dụng dữ liệu thứ cấp từ các nguồn công khai như website rao bán nhà, báo cáo thị trường hoặc cơ sở dữ liệu mở. Không bao gồm dữ liệu cá nhân hoặc thông tin nội bộ từ doanh nghiệp.

5. Phương pháp nghiên cứu

Phương pháp lý thuyết: tổng hợp, phân tích tài liệu liên quan (thị trường bất động sản và các mô hình dự đoán giá nhà; học máy, học sâu; dữ liệu chuỗi thời gian, mạng nơ - ron hồi tiếp, LSTM, v.v).

Phương pháp thực nghiệm: Triển khai các thí nghiệm trên bộ dữ liệu cụ thể, đánh giá kết quả bằng các độ đo định lượng.

Chương 1: TỔNG QUAN CƠ SỞ LÝ THUYẾT

1.1. Bài toán dự đoán giá nhà và chuỗi thời gian

1.1.1. Bài toán dự đoán giá nhà

Dự đoán giá nhà là một bài toán kinh điển trong lĩnh vực trí tuệ nhân tạo và có ý nghĩa, mang lại giá trị thực tế cao trong lĩnh vực kinh tế xã hội, đầu tư bất động sản, và công tác hoạch định các chính sách đô thị. Mục tiêu bài toán hướng tới là thiết kế một mô hình có khả năng ước lượng chính xác giá cho mỗi sản phẩm nhà đất trên cơ sở những đặc trưng, đặc biệt là có yếu tố thời gian.

1.1.2. Chuỗi thời gian

Chuỗi thời gian (time series) là 1 tập những giá trị theo dõi được sắp xếp theo trật tự thời gian như: giây, phút, giờ, ngày, tháng, v.v. Chuỗi thời gian có 4 đặc điểm chính đó là tính mùa vụ, tính vòng lặp, tính xu hướng và tính bất thường.

Quy trình dự báo gồm các bước chính như sau: chuẩn bị dữ liệu; huấn luyện mô hình; kiểm tra và so sánh; cải thiện hiệu suất; dự báo.

1.2. Một số mô hình học máy truyền thống

1.2.1. Các mô hình hồi quy cơ bản

Hồi quy tuyến tính là 1 mô hình nền tảng và quan trọng trong nhóm học máy có giám sát, thường được dùng cho các trường hợp dự báo giá trị thay đổi liên tiếp.

Mục tiêu cốt lõi của việc huấn luyện mô hình là làm giảm tổng sai lệch bình phương giữa giá trị dự đoán với giá trị thực tế, bằng cách ứng dụng các thuật toán tối ưu như giảm Gradient.

Tuy nhiên, mối liên hệ giữa giá nhà trong thực tế và những đặc trưng thường không tuân theo dạng tuyến tính đơn giản. Nhiều hiện tượng có tính phi tuyến hoặc tồn tại các tương tác phức tạp giữa các biến, khiến

cấu trúc dữ liệu không thể được nắm bắt đầy đủ bởi mô hình tuyến tính. Vì vậy, cần phải mở rộng sang những mô hình hồi quy phi tuyến, đặc biệt là mô hình hồi quy đa thức để mô hình có khả năng thể học được những mối liên hệ phi tuyến và những tương tác bậc cao giữa những đặc trưng.

1.2.2. Mô hình AutoRegressive Integrated Moving Average (ARIMA)

ARIMA là 1 mô hình kinh điển trong lĩnh vực dự báo chuỗi thời gian và đã được ứng dụng rộng rãi. ARIMA dựa trên các giá trị trong quá khứ của chuỗi để ước lượng và dự đoán kết quả ở tương lai. Mô hình này có khả năng mô tả chuỗi thời gian có xu hướng mùa vụ, nhưng cũng có thể gặp một số hạn chế đối với dữ liệu phi tuyến, hoặc những yếu tố ảnh hưởng phức tạp.

1.2.3. Mô hình rừng ngẫu nhiên (Random Forest - RF)

Rừng ngẫu nhiên là mô hình học kết hợp có thể áp dụng được cho cả 2 trường hợp hồi quy và phân loại. Mô hình này tổng hợp thành quả từ những cây quyết định để đưa ra kết quả dự đoán duy nhất, nên có tính linh hoạt và dễ sử dụng, đạt độ chuẩn xác lớn. Rừng ngẫu nhiên xây dựng 1 tập hợp các cây quyết định và mỗi cây sẽ được tiến hành huấn luyện trong 1 tập hợp con ngẫu nhiên của những đặc trưng và thay bằng phụ thuộc vào 1 cây độc nhất, RF khai thác "sức mạnh tập thể" để tạo ra những dự đoán đáng tin cậy hơn. Việc tăng lượng cây trong rừng sẽ giúp nâng cao sự chính xác và hạn chế hiện tượng quá khớp.

Trong khi đó, những mô hình truyền thống thường giả định rằng giữa các biến có quan hệ tuyến tính. Tuy nhiên, dữ liệu thực tế, nhất là dữ liệu giá nhà, thường mang tính phi tuyến, đồng thời chịu tác động bởi những yếu tố phức tạp như chính sách, tâm lý xã hội hay đặc thù địa phương. Ngoài ra, đối với những mô hình truyền thống còn yêu cầu dữ liệu cần thỏa mãn những giả định nghiêm ngặt như tính dừng, phân phối chuẩn của sai số và không có tự tương quan dư thừa, khiến việc áp dụng vào thực tiễn còn vướng nhiều khó khăn. Hiện nay, dữ

liệu ngày càng nhiều và phức tạp về cấu trúc thì tính linh hoạt cũng như khả năng mở rộng của những mô hình này dần bộc lộ rõ hạn chế.

1.3. Các mô hình dự đoán học sâu

1.3.1. Mạng Artificial Neural Networks (ANN)

Trong ANN, mỗi nơ-ron thực hiện 1 phép biến đổi tuyến tính có trọng số, sau đó dùng hàm kích hoạt nhằm tạo ra đầu ra. ANN có thể học những hàm ánh xạ phức tạp, song lại thường gặp hạn chế trong xử lý chuỗi dữ liệu có yếu tố phụ thuộc vào thời gian.

ANN có thể học các liên hệ phi tuyến giữa dữ liệu đầu vào và đầu ra, điều này không thể hiện ở những mô hình truyền thống. Ví dụ, mối quan hệ giữa vị trí địa lý, diện tích, hệ thống tiện ích xung quanh ngôi nhà với giá nhà thường không là tuyến tính nhưng ANN có thể học được những mẫu phức tạp này.

Tuy nhiên, ANN truyền thống cũng tồn tại những điểm yếu, nhất là trong giải quyết kiểu dữ liệu chuỗi thời gian. Do thiếu cơ chế ghi nhớ trạng thái trước đó, ANN thường không khai thác được những mối liên quan hệ phụ thuộc theo thời gian tại tập dữ liệu.

Tóm lại, ANN mang lại khả năng có thể học phi tuyến và xử lý dữ liệu có độ phức tạp, đây là cơ sở quan trọng trong học sâu. Dù còn có những hạn chế trong xử lý chuỗi thời gian nhưng mô hình ANN vẫn là bước đầu tiên cần thiết để phát triển những mô hình tiên tiến hơn như mô hình RNN và LSTM vốn phù hợp hơn với dự đoán giá nhà theo thời gian tại Việt Nam.

1.3.2. Mạng Convolutional Neural Networks (CNN)

CNN là 1 loại mạng nơ-ron chuyên biệt, có độ chính xác rất lớn trong giải quyết loại dữ liệu có cấu trúc dạng lưới, nhất là dữ liệu ảnh. Xuất phát từ cảm hứng về vỏ não thị giác con người, CNN có thể tự học và hoàn toàn thích nghi đối với các hệ thống phân cấp không gian

của đặc trưng của dữ liệu đầu vào. Vì vậy, CNN trở thành kiến trúc nền tảng cho phần lớn các tác vụ thị giác máy tính hiện nay.

1.3.3. Mạng Recurrent Neural Networks (RNN)

RNN là 1 loại mạng noron đặc biệt được xây dựng nhằm xử lý bộ dữ liệu theo trình tự, trong đó thông tin tại thời điểm hiện thời có thể phụ thuộc vào những trạng thái đã diễn ra trước đó. Khác những mạng noron truyền thống, mạng RNN có khả năng ghi nhớ dữ liệu trong quá khứ qua cơ chế phản hồi nội tại, từ đó tạo làm cho mô hình học được các mối liên hệ theo thời gian của chuỗi dữ liệu.

Tuy nhiên, RNN truyền thống sẽ gặp một vấn đề lớn trong thời gian huấn luyện, đó là hiện tượng bị biến mất, hoặc bị bùng nổ gradient. Khi chuỗi dữ liệu quá dài, các gradient được truyền ngược qua nhiều bước thời gian có thể trở nên quá bé hoặc quá to, làm cho mô hình khó khăn để học được những phụ thuộc dài hạn một cách hiệu quả. Hệ quả là RNN chỉ có thể ghi nhớ thông tin trong ngắn hạn, làm giảm độ chính xác đối với những bài toán cần nhận diện xu hướng dài hạn, chẳng hạn như dự đoán giá nhà.

1.3.4. Mô hình Long Short-Term Memory (LSTM)

LSTM là 1 kiến trúc cải RNN nhằm giải quyết hiện tượng biến mất gradient và các phụ thuộc dài hạn. Tương tự RNN và LSTM cũng có cấu trúc tuần tự nhưng những đơn vị lặp lại sẽ được thiết kế phức tạp hơn nên nó có thể truy xuất dữ liệu từ 1 tập dữ liệu lớn hơn rất nhiều, nhờ đó rất phù hợp trong các trường hợp dự báo có yếu tố chuỗi thời gian dài hạn.

Bên cạnh đó, một trong các ưu điểm của LSTM kể đến là nó có thể mở rộng và phối kết hợp với những kiến trúc học sâu khác. LSTM có thể được xếp chồng thành nhiều lớp để tăng độ sâu của mô hình hoặc có thể kết hợp các mạng tích chập trong xử lý dữ liệu không gian, ví dụ bản đồ vị trí bất động sản. LSTM còn có thể tích hợp với attention mechanism để tăng khả năng tập trung vào các phần quan trọng của

chuỗi dữ liệu. Ngoài ra, LSTM có một số điểm yếu như tốc độ huấn luyện chậm, cần số lượng dữ liệu đủ lớn để có khả năng nâng cao hiệu quả dự đoán.

1.4. Cơ sở dữ liệu giá nhà

1.4.1. Trên thế giới

Cơ sở dữ liệu giá nhà toàn cầu có thể truy cập từ các tổ chức uy tín như: Ngân hàng Dự trữ Liên bang Dallas, BIS, OECD, IMF, Global Property Guide và Numbeo. Các nguồn này cung cấp dữ liệu theo quốc gia, chỉ số giá, tỷ lệ thu nhập, tỷ suất lợi nhuận cho thuê và xu hướng thị trường.

1.4.2. Tại Việt Nam

Dữ liệu trong lĩnh vực bất động sản ở Việt Nam hiện nay cũng đã có những nguồn cung cấp phong phú, đa dạng và số lượng nhiều hơn trước. Tuy nhiên, dữ liệu vẫn đang ở trong tình trạng phân tán và thiếu đồng bộ. Hiện nay, thông tin về giá nhà được cung cấp qua một số nền tảng thương mại như Batdongsan.com.vn, Homedy, Guland, Landviet, v.v. cũng như từ nguồn dữ liệu mở của nhà nước, các báo cáo thị trường do một số tổ chức cung cấp.

1.5. Một số công trình nghiên cứu liên quan đến đề tài

1.5.1. Công trình nghiên cứu trong nước

Luận văn "Dự đoán giá bất động sản tại Quận 7, Thành phố Hồ Chí Minh bằng mô hình LSTM" của Đỗ Hoàng Hiệp, Đại học Quốc gia Tp Hồ Chí Minh.

Đề tài khoa học "Dự đoán giá nhà sử dụng dữ liệu Kaggle" của tác giả Vũ Đăng Hoàng Long & Nguyễn Huỳnh Đại Lợi, Huỳnh Long Nam, Trần Thanh Phúc, Nguyễn Đăng Trung Tiến - Đại học Khoa học tự nhiên Tp Hồ Chí Minh.

Bài viết "Công nghệ dự đoán giá nhà bằng Machine Learning giúp tối ưu lợi nhuận" tại HBR Việt Nam

Đồ án tốt nghiệp "Mô hình hồi quy tuyến tính và ứng dụng trong dự báo giá nhà" - Đại học Bách khoa Hà Nội của Đỗ Văn Hùng.

Khóa luận tốt nghiệp "Ứng dụng một số mô hình học máy trong dự đoán giá bất động sản tại Hà Nội" của tác giả Trịnh Thị Thu Hòa - Học viện Ngân hàng

1.5.2. Các công trình nghiên cứu quốc tế

Bài báo "House Price Prediction Using LSTM" của nhóm tác giả Xiaochen Chen, Lai Wei, Jiaxin Xu - Đại học Khoa học & Công nghệ Hồng Kông

Bài báo "Comparison of Real Estate Price Prediction Based on LSTM and LGBM" - Đại học Nông nghiệp Trung Quốc của tác giả Songyi Shi.

Bài báo "Linear Regression for House Price Prediction" - NRI Institute of Technology, Ấn Độ của tác giả V. Tejaswini và cộng sự.

Bài báo "Real Estate Price Prediction Using Deep Learning Techniques" đăng trên Springer Journal.

Bài báo "House Price Prediction Using LSTM Neural Network" - IEEE Conference Proceedings của nhiều nhóm nghiên cứu quốc tế.

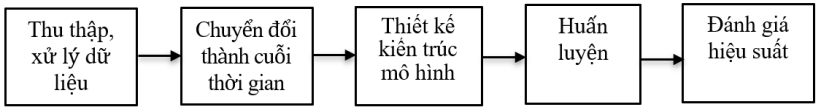
Chương 2: THIẾT KẾ VÀ THỰC NGHIỆM MÔ HÌNH DỰ ĐOÁN GIÁ NHÀ BẰNG LONG SHORT-TERM MEMORY

2.1. Mô hình LSTM trong dự đoán giá nhà

Khả năng học và nắm bắt những xu hướng dài hạn trong xử lý dữ liệu về giá nhà là ưu điểm nổi bật của LSTM. Thị trường nhà đất vốn mang tính chu kỳ và chịu sự tác động bởi nhiều yếu tố như: Những chính sách tín dụng, tốc độ đô thị hóa, mức lãi suất ngân hàng và tâm lý người dùng. Nó không chỉ tác động ngay lập tức mà còn để lại tác động kéo dài theo thời gian.

Với những khả năng học phụ thuộc dài hạn, xử lý đối với dữ liệu phi tuyến và tích hợp đa đặc trưng, LSTM được xem là một công cụ vượt trội, rất phù hợp để dùng trong bài toán dự đoán giá nhà đất tại Việt Nam. Mô hình LSTM mở ra hướng tiếp cận mới cho những nghiên cứu và đẩy mạnh ứng dụng trong lĩnh vực dự đoán giá trị bất động sản để hỗ trợ ra các quyết định đầu tư, hoạch định các cơ chế, chính sách và quản lý thị trường được hiệu quả hơn.

Quy trình xây dựng mô hình gồm các bước như Hình 2.1:



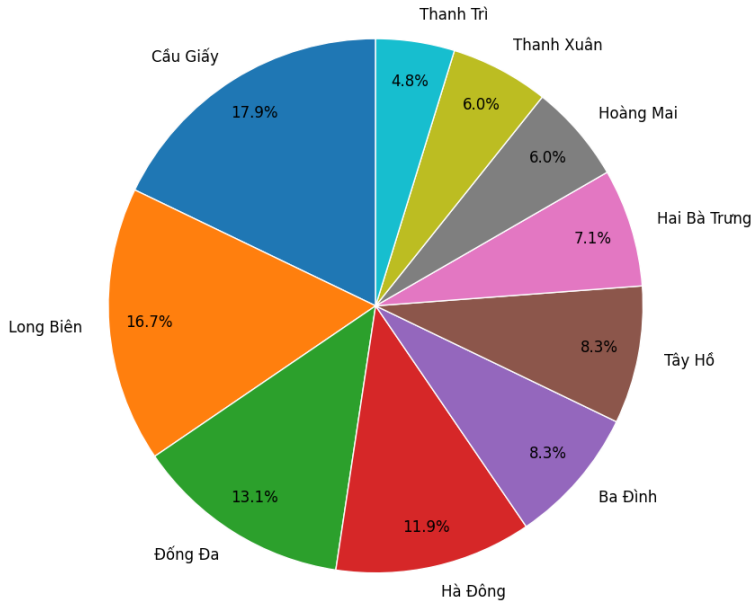
Hình 2.1: Quy trình xây dựng mô hình dự đoán giá nhà bằng LSTM

2.2. Thu thập và xử lý dữ liệu

Quá trình thu thập dữ liệu giá nhà được tiến hành bằng kỹ thuật cào dữ liệu tự động, trong đó Selenium được dùng để điều khiển trình duyệt và tải các trang web mục tiêu.

Phạm vi thu thập URL được giới hạn trong mục "Nhà riêng" thuộc khu vực Hà Nội trên trang *batdongsan.com*. Các bài đăng được phân trang, với mỗi trang chứa tối đa 20 kết quả. Quy trình cào dữ liệu tập trung vào 30 trang đầu tiên, tương ứng với khoảng từ 500 đến

600 URL ban đầu. Sau quy trình thu thập và làm sạch, ta thu được bộ dữ liệu với tỷ lệ tương ứng trong Hình 2.4:



Hình 2.4: Biểu đồ Top 10 quận/huyện (cũ) của Hà Nội có số lượng tin đăng mua bán nhà riêng cao nhất giai đoạn từ tháng 11/2024 đến tháng 10/2025

2.3. Xây dựng tập dữ liệu chuỗi thời gian

Quá trình hình thành tập dữ liệu kiểu chuỗi thời gian trong đoạn mã trên được thực hiện bằng cách tổ chức và chuẩn hóa thông tin giá nhà theo các mốc thời gian cụ thể.

Mỗi bản ghi trong chuỗi thời gian được tổ chức thành 3 trường: mã bài đăng (post_id), thời gian (time) và giá/m² (price_per_square_meter. Điều này giúp dữ liệu không chỉ mang tính đại diện cho từng bài đăng cụ thể mà còn phản ánh xu hướng giá chung của khu vực.

Sau đó, dữ liệu được sắp xếp tăng dần của post_id và time, đảm bảo rằng các thay đổi giá của cùng một bất động sản được xuất hiện theo trình tự thời gian.

Quá trình chuẩn hóa giúp hạn chế tác động của những giá trị bất thường (outliers) và làm cho cơ sở dữ liệu trở nên đồng nhất hơn, đặc biệt trong trường hợp các thuộc tính có vùng giá trị khác nhau.

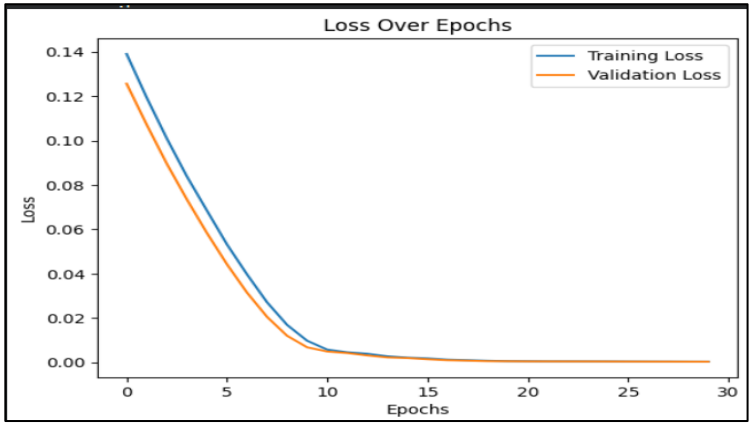
Sau khi xử lý, bộ dữ liệu được chia thành 2 phần: 80% dành cho huấn luyện, 20% dành để kiểm tra, nhằm đảm bảo khách quan đối với việc đánh giá kết quả hoạt động của mô hình.

2.4. Thiết kế kiến trúc mô hình LSTM

Mô hình đã được thiết kế với kiến trúc đơn giản nhưng thực sự hiệu quả, gồm 1 lớp LSTM đầu vào, 1 lớp Dense đầu ra. Lớp đầu vào tiếp nhận chuỗi dữ liệu có độ dài sequence_length = 6, tức là sử dụng 6 tháng trước đấy để dự đoán giá tại thời gian tiếp theo. Mỗi một bước thời gian có 1 đặc trưng duy nhất là giá đã được chuẩn hóa (price_scaled). Do đó, shape đầu vào của mô hình là (6, 1). Lớp LSTM gồm 50 unit, dùng hàm ReLU để làm tăng tính phi tuyến, đồng thời giúp mô hình có thể nhận diện những mẫu phức tạp của chuỗi dữ liệu. Lớp này chịu trách nhiệm ghi nhớ, xử lý thông tin theo thời gian. Lớp đầu ra gồm một noron duy nhất, không sử dụng hàm kích hoạt (hoặc hàm kích hoạt tuyến tính ngầm định), dùng để dự đoán những giá trị liên tục, tức là giá của bất động sản ở thời gian tiếp theo.

2.5. Huấn luyện và đánh giá mô hình

Biểu đồ Loss Over Epochs tại Hình 2.5 cung cấp các thông tin quan trọng trong quá trình huấn luyện mô hình LSTM, trong đó hàm mất mát được dùng là MSE. MSE chính là lựa chọn phổ biến, rất phù hợp đối với bài toán hồi quy, đặc biệt khi mục tiêu là dự đoán 1 giá trị liên tục.



Hình 2.5: Biểu đồ Loss Over Epochs

Trong thời gian đầu huấn luyện thì cả Training Loss và Validation Loss đều giảm mạnh, chứng tỏ mô hình học nhanh từ bộ dữ liệu, đồng thời nắm bắt được những mẫu phức tạp trong chuỗi thời gian. Sau khoảng 10 epoch, cả hai đường loss dần ổn định ở mức rất thấp và không có sự dao động đáng kể, điều này chứng tỏ mô hình đã đạt trạng thái hội tụ. Kết quả này khẳng định mô hình không những học tốt mà còn duy trì được tính ổn định trong suốt quá trình huấn luyện.

2.6. Nhận xét, đánh giá chung

Mô hình được xây dựng và thực nghiệm trên dữ liệu giá nhà đất tại Hà Nội từ tháng 11/2024 đến tháng 10/2025. Các khu vực được chọn bao gồm cả quận trung tâm và vùng ven, nhằm kiểm tra tính khái quát của mô hình. Kết quả thực nghiệm đã chứng minh mô hình LSTM có khả năng dự báo chính xác các xu hướng tăng/giảm giá theo thời gian, nhất là tại những vùng có tập dữ liệu ổn định và đầy đủ.

Mặc dù vậy, mô hình vẫn còn những yếu điểm nhất định như huấn luyện mô hình cần số lượng dữ liệu lớn, chất lượng cao và nguồn tài nguyên tính toán mạnh. Ngoài ra, mô hình mang tính "hộp đen", khiến việc giải thích kết quả cho người dùng cuối trở nên khó khăn.

Chương 3. KẾT QUẢ THỰC NGHIỆM VÀ LUẬN BÀN

3.1. Kết quả thực nghiệm

Sau khi hoàn thành xong toàn bộ quá trình thiết kế và huấn luyện mô hình LSTM trên tập dữ liệu giá nhà thực tế tại Việt Nam, mô hình đã được tiến hành thử nghiệm tại Hà Nội trong khoảng từ tháng 11/2024 đến tháng 10/2025. Dữ liệu đã được chuẩn hóa, chuyển đổi thành chuỗi thời gian theo từng quận, huyện (cũ) với mỗi điểm dữ liệu đại diện cho giá nhà trung bình theo tháng.

Bảng 3.1: Kết quả của mô hình LSTM đánh giá trên bộ dữ liệu kiểm thử (Testing dataset)

Độ đo	Kết quả
MAE	9.18 (triệu/m^2)
MSE	161.11
RMSE	12.69
Coefficient of Determination (R^2)	0.9861

So sánh hiệu hiệu quả của các mô hình

Để xác định được hiệu quả của mô hình LSTM, nghiên cứu đã so sánh với những mô hình như ARIMA và hồi quy tuyến tính. Tất cả những mô hình này đều được huấn luyện trên cùng 1 bộ dữ liệu và cùng được đánh giá bằng các chỉ số tương tự.

Kết quả tại Bảng 3.3 đánh giá hiệu quả của các mô hình LSTM, Linear Regression, ARIMA đối với bài toán dự đoán giá nhà theo chuỗi thời gian cho thấy sự khác biệt rõ ràng giữa các công cụ.

Bảng 3.3: So sánh hiệu suất của các mô hình

Mô hình	MAE	MSE	RMSE	R^2
Linear Regression	7.72	113.55	10.66	0.9902
LSTM	9.18	161.11	12.69	0.9861
ARIMA	117.14	28064.14	117.14	0.15

Mô hình Linear Regression đạt hiệu suất rất tốt với MAE = 7.72, MSE = 113.55, RMSE = 10.66 và R^2 = 0.9902. Kết quả này chứng tỏ rằng mô hình tuyến tính có khả năng giải thích gần như toàn bộ biến động trong dữ liệu. Tuy nhiên, Linear Regression không thể xử lý các phụ thuộc phi tuyến hoặc xu hướng thời gian phức tạp, khiến nó hạn chế trong việc nắm bắt những biến động đột ngột trong chuỗi dữ liệu.

LSTM cũng đã đạt hiệu suất ổn định với R^2 = 0.9861, nhưng sai số tuyệt đối (MAE = 9.18, MSE = 161.11, RMSE = 12.69) cao hơn so với Linear Regression. Điều này có thể do LSTM cố gắng học các mẫu phi tuyến và phụ thuộc thời gian phức tạp hơn - điều mà mô hình tuyến tính sẽ bỏ qua. Đối với trường hợp dự đoán giá nhà theo chuỗi thời gian, LSTM có ưu thế đặc biệt nhờ khả năng nắm bắt cả xu hướng dài hạn lẫn ngắn hạn, đồng thời nhận diện được những biến động bất thường trong dữ liệu.

Ngược lại, mô hình ARIMA hoàn toàn thất bại với R^2 = 0.15, MAE = 117.14, MSE = 28064.14 và RMSE = 117.14. Kết quả cho thấy ARIMA không phù hợp cho dự đoán giá nhà theo chuỗi thời gian.

Tóm lại, Linear Regression là lựa chọn hiệu quả nếu dữ liệu có cấu trúc tuyến tính rõ ràng, trong khi LSTM phù hợp hơn khi cần xử lý các phụ thuộc phi tuyến và xu hướng thời gian phức tạp. ARIMA nên được loại bỏ đối với bài toán này.

3.2. Đánh giá kết quả

3.2.1. Hiệu quả mô hình

Mô hình LSTM đã chứng minh được khả năng học các phụ thuộc dài hạn trong chuỗi thời gian, giải quyết hiệu quả đối với dữ liệu phi tuyến, đa chiều, đồng thời đáp ứng tốt với biến động thị trường. Việc đạt được sai số thấp và hệ số tương quan cao cho thấy mô hình hoàn toàn có thể dùng như một công cụ hỗ trợ để ra quyết định trong lĩnh vực bất động sản.

3.2.2. Ý nghĩa thực tiễn

Hiện nay, thị trường bất động sản tại Việt Nam đang từng bước chuyển đổi số, việc áp dụng mô hình LSTM mang đến nhiều giá thực tiễn. Đối với nhà đầu tư, mô hình có khả năng giúp hỗ trợ dự đoán chính xác hơn xu thế biến động giá bất động sản để đưa ra các quyết định đầu đúng đắn.

Với người mua nhà, đặc biệt là lần đầu tiếp cận thị trường, mô hình cung cấp công cụ hỗ trợ đáng tin cậy để đánh giá mức giá hợp lý của giá nhà tại một số khu vực cụ thể. Qua đó, mô hình có khả năng hỗ trợ người mua lựa chọn căn nhà phù hợp so với tầm tài chính và nhu cầu cá nhân.

Đối với cơ quan quản lý, mô hình có vai trò như 1 công cụ hữu ích trong theo dõi sự xu hướng trong thị trường bất động sản. Từ kết quả dự đoán giúp đánh giá hiệu quả các chính sách hiện hành, kịp thời điều chỉnh cho phù hợp. Ngoài ra, dữ liệu từ mô hình cũng hỗ trợ quá trình hoạch định quy hoạch đô thị, giúp phân bổ nguồn lực hợp lý, đồng thời đáp ứng tốt hơn nhu cầu người dân trong thời gian tới.

3.2.3. Hạn chế

Mô hình đã đạt được những kết quả đáng ghi nhận, tuy nhiên mô hình vẫn có một số hạn chế cần khắc phục. Những hạn chế này

xuất phát từ cả chất lượng của dữ liệu đưa vào, phương pháp dự đoán cũng như khả năng giải thích kết quả.

Nguồn dữ liệu thu thập được còn khá giới hạn cả về phạm vi địa lý lẫn loại hình bất động sản. Hiện tại, mô hình chỉ tập trung vào nhà riêng tại khu vực Hà Nội, trong khi chưa mở rộng sang những sản phẩm bất động sản khác, ví dụ như căn chung cư, biệt thự hay đất nền. Hơn nữa, sự phân bố dữ liệu không đồng đều và có sự thiên lệch rõ rệt về một số quận huyện cụ thể, dẫn đến nguy cơ tạo ra kết quả dự đoán thiếu tính đại diện. Bên cạnh đó, quy mô bộ dữ liệu còn khá khiêm tốn với chỉ hơn 100 mẫu. Với số lượng mẫu ít, mô hình có thể gặp khó khăn trong việc học các quy luật phức tạp hoặc đảm bảo độ tin cậy khi ứng dụng vào thực tế.

Về phương pháp, mô hình dự đoán được xây dựng chủ yếu dựa vào giá trị lịch sử (như dữ liệu giá từ các tháng trước) làm yếu tố đầu vào duy nhất để đưa ra dự báo. Tuy nhiên, còn nhiều yếu tố ảnh hưởng không nhỏ đến giá trị nhà đất nhưng chưa được tận dụng do thiếu dữ liệu. Ví dụ, một căn nhà nằm ở quận Hoàn Kiếm (gần trung tâm Hà Nội) thường có giá trị cao hơn đáng kể so với một căn nhà cùng diện tích ở vùng ven Hà Nội, chẳng hạn như huyện Ba Vì. Hay như một căn nhà gần trường học, bệnh viện hoặc công viên sẽ hấp dẫn hơn và có giá trị lớn hơn so với một căn nhà không có các tiện ích này. Tương tự, những yếu tố kinh tế vĩ mô như lãi suất ngân hàng tăng cao có thể khiến người mua hạn chế vay vốn, hoặc chính sách siết chặt cấp giấy phép xây dựng có thể làm giảm nguồn cung trên thị trường, qua đó tác động mạnh đến giá cả. Tuy nhiên, những thông tin về vị trí chi tiết, hạ tầng xung quanh hay biến động kinh tế vĩ mô thường không có đầy đủ trong dữ liệu hiện có. Điều này khiến mô hình không thể khai thác triệt để các yếu tố quan trọng để đưa ra kết quả dự báo chính xác. Như vậy, hạn chế không bắt nguồn từ năng lực của mô hình mà xuất phát từ việc dữ liệu đầu vào còn thiếu sót và chưa toàn diện.

Một hạn chế quan trọng khác là bản chất "hộp đen" của mô hình. Bản chất "hộp đen" của mô hình LSTM (Long Short-Term Memory) bắt nguồn từ cấu trúc, cơ chế hoạt động phức tạp của mạng nơ-ron sâu, khiến việc giải thích chi tiết quá trình xử lý và đưa ra quyết định gặp khó khăn.

3.2.4. Khả năng mở rộng

Để mở rộng khả năng áp dụng vào thực tế, mô hình có thể phát triển theo một số hướng sau:

1. Mở rộng phạm vi và quy mô dữ liệu

Hiện tại, dữ liệu chủ yếu tập trung vào nhà riêng tại khu vực Hà Nội, với số lượng mẫu hạn chế. Trong tương lai, phạm vi thu thập dữ liệu có thể mở rộng sang những loại bất động sản khác như: biệt thự, căn chung cư, đất nền, hoặc thậm chí là các loại tài sản thương mại. Đồng thời, việc mở rộng địa lý ra các tỉnh thành khác sẽ giúp mô hình trở nên đa dạng và có tính đại diện cao hơn. Ngoài ra, việc gia tăng số lượng mẫu dữ liệu bằng các phương pháp thu thập tự động từ nhiều trang bất động sản khác nhau sẽ giúp nâng cao độ chính xác và độ tin cậy của mô hình.

2. Tận dụng tối đa các yếu tố đầu vào

Mặc dù thông tin ban đầu về bất động sản cung cấp một số yếu tố khách quan trọng như vị trí địa lý, hệ thống tiện ích xung quanh, diện tích, số phòng ngủ, dữ liệu này vẫn chưa được khai thác hiệu quả do tình trạng thiếu hụt và không chuẩn hóa. Một số bài đăng chỉ cung cấp thông tin cơ bản như diện tích và vị trí, trong khi các thông tin chi tiết khác như số phòng ngủ, số phòng vệ sinh, hướng nhà hay diện tích mặt đường thường bị bỏ sót hoặc không đầy đủ. Điều này khiến mô hình có thể gặp khó khăn trong việc hiểu được cấu trúc và giá trị thực tế của thị trường. Bên cạnh đó, thông tin về vị trí địa lý thường mang tính quảng cáo, tập trung vào các yếu tố tích cực như "gần bệnh viện, trường học X" mà không đề cập đến những yếu tố tiêu cực có thể ảnh

hưởng đến giá trị tài sản, ví dụ như khu vực dính quy hoạch, tranh chấp pháp lý hoặc gần các nguồn gây ô nhiễm. Việc thiếu thông tin tiêu cực này làm giảm khả năng mô hình nắm bắt bức tranh toàn diện của thị trường bất động sản. Ngoài ra, dữ liệu thu thập được từ các bài đăng thường không được chuẩn hóa, có thể dẫn đến sự không đồng nhất trong cách biểu diễn dữ liệu. Bên cạnh việc khai thác dữ liệu về vị trí căn nhà, việc tích hợp những yếu tố như lãi suất ngân hàng, tốc độ tăng trưởng GDP hay các chính sách nhà đất mới ban hành sẽ giúp mô hình nắm bắt sát hơn biến động của thị trường.

3. Cải thiện khả năng giải thích của mô hình

Do bản chất "hộp đen" của LSTM, việc giải thích kết quả dự đoán hiện tại còn gặp khá nhiều khó khăn. Vì vậy, để giải quyết điều này, có thể ứng dụng các kỹ thuật hỗ trợ giải thích như SHAP (SHapley Additive exPlanations) hoặc phân tích độ nhạy (sensitivity analysis). Các công cụ này có thể giúp làm rõ các mức độ ảnh hưởng của mỗi yếu tố đầu vào đến kết quả cuối cùng, qua đó tăng cường niềm tin của người dùng vào mô hình.

4. Phát triển ứng dụng thực tế và tích hợp hệ thống

Mô hình có thể được tích hợp vào những nền tảng trực tuyến, hoặc các ứng dụng di động hỗ trợ người dùng trong ra quyết định liên quan đến bất động sản. Ví dụ như xây dựng một hệ thống gợi ý giá nhà dựa trên các tiêu chí đầu vào mà người dùng cung cấp hoặc phát triển công cụ để phân tích xu hướng thị trường để phục vụ nhà đầu tư và các cơ quan quản lý.

KẾT LUẬN

Kết quả của nghiên cứu

Trong đề tài ***"Áp dụng mô hình Long Short-Term Memory trong dự đoán giá nhà tại Việt Nam"***, mô hình LSTM đã được nghiên cứu, thiết kế và triển khai nhằm giải quyết bài toán dự đoán giá nhà tại Việt Nam theo chuỗi thời gian. Thông qua quá trình xây dựng kiến trúc mô hình, xử lý dữ liệu thực tế và thực nghiệm tại khu vực Hà Nội, nghiên cứu đã chứng minh rằng LSTM là một công cụ hiệu quả trong việc khai thác các mối quan hệ dài hạn và phi tuyến trong dữ liệu bất động sản. Kết quả thực nghiệm cho thấy mô hình LSTM có khả năng dự đoán chính xác xu hướng giá nhà, phản ứng tốt với các biến động thị trường và vượt trội so với các mô hình truyền thống như hồi quy tuyến tính và ARIMA. Mô hình không chỉ học được các đặc điểm nội tại của chuỗi thời gian mà còn có thể tích hợp các đặc trưng bổ sung như vị trí địa lý, loại hình bất động sản và các yếu tố kinh tế xã hội.

Tuy nhiên, nghiên cứu cũng nhận diện một số thách thức trong việc áp dụng LSTM vào thực tế, bao gồm yêu cầu về dữ liệu lớn và chất lượng cao, thời gian huấn luyện dài và tính khó giải thích của mô hình. Những hạn chế này đặt ra yêu cầu tiếp tục cải tiến mô hình và mở rộng hướng nghiên cứu trong tương lai.

Kiến nghị đề xuất

Dựa trên kết quả nghiên cứu và thực nghiệm, một số kiến nghị được đề xuất nhằm nâng cao hiệu quả ứng dụng mô hình LSTM trong dự đoán giá nhà tại Việt Nam trong tương lai như sau:

Về dữ liệu

Cần mở rộng phạm vi thu thập để bổ sung thêm các loại hình bất động sản khác ngoài nhà riêng, chẳng hạn như căn hộ chung cư, biệt thự, đất nền hoặc tài sản thương mại. Hiện tại, dữ liệu chủ yếu tập trung ở khu vực Hà Nội, dẫn đến sự thiên lệch về mặt địa lý. Việc mở

rộng phạm vi địa lý ra các tỉnh thành khác sẽ giúp mô hình trở nên đại diện và chính xác hơn. Đồng thời, mặc dù đã có sẵn các yếu tố như vị trí địa lý, diện tích, số phòng, tiện ích xung quanh và các yếu tố kinh tế vĩ mô, nhưng chúng chưa được khai thác triệt để. Do đó, cần phân tích sâu hơn để tích hợp hiệu quả các yếu tố này vào mô hình, chẳng hạn bằng cách áp dụng kỹ thuật tạo biến tương tác (interaction features) giữa các yếu tố để tìm ra mối quan hệ tiềm ẩn. Ngoài ra, thị trường bất động sản luôn biến động, vì vậy cần xây dựng cơ chế tự động cập nhật dữ liệu thường xuyên thông qua các phương pháp như web scraping hoặc hợp tác với các nền tảng bất động sản để lấy dữ liệu giao dịch gần đây.

Về mô hình

Nên thử nghiệm các mô hình hybrid kết hợp LSTM với các thuật toán học máy khác như XGBoost, Random Forest hoặc Gradient Boosting để tận dụng ưu điểm của từng phương pháp. Điều này không chỉ giúp cải thiện độ chính xác mà còn tăng cường khả năng giải thích kết quả, đặc biệt trong bối cảnh bản chất "hộp đen" của LSTM gây khó khăn cho người dùng không chuyên.

Về ứng dụng thực tiễn

Cần phát triển các công cụ hỗ trợ ra quyết định dành cho nhà đầu tư, người mua nhà và cơ quan quản lý. Các công cụ này cần cung cấp thông tin chi tiết về xu hướng biến động giá theo thời gian và các yếu tố ảnh hưởng, giúp người dùng đưa ra quyết định chính xác hơn. Mô hình cũng có thể được tích hợp vào các trang web hoặc ứng dụng bất động sản phổ biến, giúp cung cấp ước lượng giá trị bất động sản theo thời gian thực. Điều này không chỉ mang lại lợi ích lớn cho người dùng cuối mà còn tăng tính cạnh tranh cho các nền tảng kinh doanh bất động sản. Giao diện của các công cụ này cần thân thiện, đặc biệt đối với người dùng không chuyên, những người cần hiểu rõ cơ sở của các dự đoán trước khi đưa ra quyết định.

Hướng nghiên cứu trong tương lai

Cần nghiên cứu việc tích hợp LSTM với các kỹ thuật tiên tiến như Attention Mechanism hoặc kiến trúc Transformer để nâng cao khả năng xử lý chuỗi thời gian và dữ liệu phi cấu trúc. Ngoài việc dự đoán giá nhà, mô hình có thể được mở rộng để giải quyết các bài toán khác trong lĩnh vực bất động sản, chẳng hạn như phân tích xu hướng khu vực hoặc đánh giá rủi ro đầu tư. Với sự phát triển mạnh mẽ của thị trường bất động sản tại Việt Nam, việc áp dụng và cải tiến liên tục mô hình LSTM sẽ không chỉ mang lại giá trị học thuật mà còn góp phần hiện đại hóa ngành này thông qua chuyển đổi số. Những công cụ dự đoán thông minh này hứa hẹn sẽ trở thành một phần không thể thiếu trong tương lai, giúp các bên liên quan đưa ra quyết định chính xác và hiệu quả hơn./.