**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA KẾ TOÁN – KIỂM TOÁN**

---------\*\*\*---------



**BÁO CÁO MÔN HỌC**

**KHAI PHÁ VÀ PHÂN TÍCH DỮ LIỆU LỚN**

**ĐỀ TÀI**

**ĐỊNH GIÁ BẤT ĐỘNG SẢN DỰA TRÊN**

**MÔ HÌNH HỒI QUY**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn** | **TS.Trần Hùng Cường** |
| **Lớp** | **20232IT6088001** |
| **Sinh viên thực hiện** | **Nguyễn Lệ Hằng** |

**Hà Nội, 5/2024**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA KẾ TOÁN – KIỂM TOÁN**

---------\*\*\*---------



**BÁO CÁO MÔN HỌC**

**KHAI PHÁ VÀ PHÂN TÍCH DỮ LIỆU LỚN**

**ĐỀ TÀI**

**ĐỊNH GIÁ BẤT ĐỘNG SẢN DỰA TRÊN**

**MÔ HÌNH HỒI QUY**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn** | **TS.Trần Hùng Cường** |
| **Sinh viên thực hiện**  Nguyễn Lệ Hằng | **MSV**  2021606898 |
|  | |  |

**Hà Nội, 5/2024**

# LỜI CẢM ƠN

Để hoàn thành được đề án “Định giá bất động sản dựa trên mô hình hồi quy ”, em xin chân thành cảm ơn Ban Giám hiệu, các khoa, phòng và quý thầy, cô của trường Đại học Công Nghiệp Hà Nội, những người đã tận tình giúp đỡ và tạo điều kiện cho em trong quá trình học tập. Đặc biệt, em xin gửi lời cảm ơn sâu sắc đến thầy Trần Hùng Cường- người đã trực tiếp giảng dạy và hướng dẫn em thực hiện đề án “Định giá bất động sản dựa trên mô hình hồi quy ” này bằng tất cả lòng nhiệt tình và sự quan tâm sâu sắc.

Trong quá trình thực hiện bài đề án này, do hiểu biết còn nhiều hạn chế nên bài làm khó tránh khỏi những thiếu sót. Em rất mong nhận được những lời góp ý của quý thầy cô để bài báo cáo ngày càng hoàn thiện hơn.

Em xin chân thành cảm ơn!

# MỤC LỤC

[LỜI CẢM ƠN i](#_Toc170072197)

[MỤC LỤC ii](#_Toc170072198)

[DANH MỤC HÌNH ẢNH iv](#_Toc170072199)

[DANH MỤC BẢNG BIỂU v](#_Toc170072200)

[LỜI MỞ ĐẦU 1](#_Toc170072201)

[1. Lý do lựa chọn đề tài 1](#_Toc170072202)

[2. Nội dung phân tích 2](#_Toc170072203)

[3. Mục tiêu của đề tài 2](#_Toc170072204)

[4. Bố cục bài báo cáo 3](#_Toc170072205)

[CHƯƠNG 1 - TỔNG QUAN VỀ DỮ LIỆU LỚN 4](#_Toc170072206)

[1.1. Khái niệm dữ liệu lớn 4](#_Toc170072207)

[1.2. Một số phương pháp phân tích dữ liệu lớn 4](#_Toc170072208)

[1.3. Ứng dụng của phân tích dữ liệu lớn 5](#_Toc170072209)

[1.4. Những thách thức trong phân tích dữ liệu lớn 6](#_Toc170072210)

[1.5. Các bước cơ bản trong phân tích dữ liệu lớn 6](#_Toc170072211)

[CHƯƠNG 2 - CƠ SỞ LÝ THUYẾT VÀ MÔ HÌNH DỰ BÁO 8](#_Toc170072212)

[2.1. Tổng quan về bất động sản 8](#_Toc170072213)

[2.1.1. Khái niệm 8](#_Toc170072214)

[2.1.2. Phân loại bất động sản 8](#_Toc170072215)

[2.1.3. Đặc điểm của bất động sản 8](#_Toc170072216)

[2.1.4. Các yếu tố ảnh hưởng đến giá bất động sản 8](#_Toc170072217)

[2.2. Mô hình dự báo 9](#_Toc170072218)

[2.2.1. Hồi quy tuyến tính: 9](#_Toc170072219)

[2.2.2. Hồi quy đa thức 10](#_Toc170072220)

[2.2.3. Đánh giá mô hình: 10](#_Toc170072221)

[2.3. Tóm tắt các nghiên cứu trước 11](#_Toc170072222)

[CHƯƠNG 3 - MÔ HÌNH ĐỊNH GIÁ BẤT ĐỘNG SẢN 12](#_Toc170072223)

[3.1. Giới thiệu dữ liệu 12](#_Toc170072224)

[3.1.1. Tập dữ liệu 12](#_Toc170072225)

[3.1.2. Làm sạch dữ liệu 12](#_Toc170072226)

[3.1.3. Biến đổi dữ liệu 14](#_Toc170072227)

[3.1.4. Kiểm tra và loại bỏ đa cộng tuyến (multicollinearity) 14](#_Toc170072228)

[3.2. Khám phá dữ liệu (Exploratory Data Analysis - EDA) 15](#_Toc170072229)

[3.2.1. Ma trận Tương quan với Giá Nhà 16](#_Toc170072230)

[3.2.2. Mối Quan Hệ Giữa Các Biến Số Khác 16](#_Toc170072231)

[3.3. Xây dựng mô hình hồi quy 21](#_Toc170072232)

[3.3.1. Mô hình hồi quy tuyến tính 21](#_Toc170072233)

[3.3.2. Mô hình hồi quy đa thức: 22](#_Toc170072234)

[3.4. Phân tích kết quả hồi quy 23](#_Toc170072235)

[3.4.1. Phân tích kết quả mô hình hồi quy tuyến tính 23](#_Toc170072236)

[3.4.2. Phân tích kết quả mô hình hồi quy đa thức 27](#_Toc170072237)

[3.4.3. So sánh kết quả hai mô hình 32](#_Toc170072238)

[3.5. Ứng dụng kết quả mô hình hồi quy tuyến tính 37](#_Toc170072239)

[KẾT LUẬN VÀ KIẾN NGHỊ 39](#_Toc170072240)

[TÀI LIỆU THAM KHẢO 40](#_Toc170072241)

# DANH MỤC HÌNH ẢNH

[Hình 3.1. Biểu đồ Box Plot để xác định giá trị ngoại lai 13](#_Toc170071867)

[Hình 3.2 Ma trận Tương quan với Giá nhà 16](file:///C:\Users\Admin\Documents\Ki_6\KP&PTDLL1\Bao_cao_KPDLL.docx#_Toc170071868)

[Hình 3.3 Biểu đồ thể hiện mối quan hệ giữa giá nhà và ngày giao dịch 17](#_Toc170071869)

[Hình 3.4 Biểu đồ thể hiện mối quan hệ giữa giá nhà và khoảng cách tới ga MRT 18](file:///C:\Users\Admin\Documents\Ki_6\KP&PTDLL1\Bao_cao_KPDLL.docx#_Toc170071870)

[Hình 3.5 Biểu đồ mối quan hệ giữa giá nhà và vĩ độ 19](file:///C:\Users\Admin\Documents\Ki_6\KP&PTDLL1\Bao_cao_KPDLL.docx#_Toc170071871)

[Hình 3.6 Biểu đồ phân bố giá nhà trên mỗi đơn vị diện tích 20](#_Toc170071872)

[Hình 3.7 Biểu đồ Phân Tán Giữa Giá Trị Thực Tế và Giá Trị Dự Đoán 24](#_Toc170071873)

[Hình 3.8 Biểu đồ phân phối phần dư 25](#_Toc170071874)

[Hình 3.9 Biểu đồ phân tích phần dư so với giá trị thực tế 26](#_Toc170071875)

[Hình 3.10 Biểu Đồ Phân Tán Giữa Giá Trị Thực Tế và Giá Trị Dự Đoán 29](#_Toc170071876)

[Hình 3.11 Biểu đồ phân phối phần dư 30](#_Toc170071877)

[Hình 3.12 Đồ thị phần dư so với giá trị thực tế 31](#_Toc170071878)

# DANH MỤC BẢNG BIỂU

[Bảng 3.1. Xử lý dữ liệu thiếu 13](#_Toc170071879)

[Bảng 3.2. Bảng hiển thị VIF cho từng biến số 15](#_Toc170071880)

[Bảng 3.3. Các hệ số của mô hình 21](#_Toc170071881)

[Bảng 3.4. Mô hình hồi quy đa thức 22](#_Toc170071882)

[Bảng 3.5. Bảng so sánh kết quả hai mô hình 32](#_Toc170071883)

# LỜI MỞ ĐẦU

## Lý do lựa chọn đề tài

Trong thời kỳ chuyển đổi mạnh mẽ của công nghệ. Định giá bất động sản là vấn đề quan trọng và thiết yếu đối với các nhà đầu tư, doanh nghiệp bất động sản, cơ quan quản lý. Kết quả nghiên cứu giúp cung cấp những công cụ hữu ích, hỗ trợ trực tiếp các bên liên quan trong việc định giá và ra quyết định đầu tư một cách hiệu quả và chính xác

Chính ở đây, trong tình thế đầy thách thức này, đề tài “định giá bất động sản dựa trên mô hình hồi quy” trở nên cực kỳ cấp thiết. Mô hình này không chỉ giúp mô phỏng một cách chính xác hơn sự biến động của giá cả, mà còn mang lại một cơ hội quý báu để hiểu rõ hơn về yếu tố ảnh hưởng đến giá cả, từ đặc điểm kỹ thuật đến sự biến động của thị trường.

Với lượng dữ liệu ngày càng lớn và phức tạp, Machine Learning đem lại một cơ hội hiếm có để xử lý và hiểu sâu về các mối liên hệ phức tạp giữa các biến số. Sử dụng mô hình hồi quy để nhận diện các xu hướng không rõ ràng và mô hình hóa chúng không chỉ giúp dự đoán giá cách chính xác mà còn mở ra cánh cửa cho những hiểu biết mới về thị trường bất động sản.

Đề tài có khả năng cập nhật,phân tích sâu và thích ứng rất cao. Rất phù hợp với thị trường bất động sản luôn biến động rất cần mô hình định giá linh hoạt và có khả năng cập nhật tốt. Và mô hình hồi quy có tính ưu việt hơn so với các phương pháp truyền thống, cho phép chúng ta dễ dàng bổ sung, hiệu chỉnh các biến số để phản ánh chính xác tình hình thị trường. Không chỉ vậy, mô hình hồi quy cho phép xác định ảnh hưởng của các yếu tố như vị trí, diện tích, phòng ngủ…vv đến giá bất động sản. Từ những yếu tố đó giúp chúng ta hiểu sâu hơn về cơ chế định giá và các nhân tố then chốt.

Hơn thế, đề tài mang tính khóa, phổ biến rất cao và khả năng tích hợp công nghệ hiện đại đang là những vấn đề hot trong thời đại ngày nay.Việc tích hợp các công nghệ mới như Big Data, Machine Learning giúp chúng ta nâng cao được độ chính xác và tính từ động hóa của mô hình định giá. Kết hợp với phương pháp hồi quy là một công cụ phân tích định lượng được áp dụng rộng rãi trong nhiều lĩnh vực. Các kết quả nghiên cứu có thể so sánh, tham khảo với các nghiên cứu tương tự ở trong và ngoài nước. Để từ đó, giúp cho mô hình trở nên hiệu quat hơn trong thực tiễn ứng dụng.

Đồng thời, đề tài này không chỉ đặt ra những thách thức trong việc thu thập và xử lý dữ liệu mà còn đòi hỏi sự sáng tạo trong việc áp dụng các mô hình hồi quy và kỹ thuật phân tích. Thông qua những nỗ lực này, em hy vọng mở ra một hướng nghiên cứu mới và đem lại giá trị thực tế cho cả cộng đồng bất động sản.

Nhìn chung, tính cấp thiết của đề án này không chỉ nằm ở khía cạnh nghiên cứu khoa học mà còn nằm ở khả năng áp dụng và đóng góp vào thực tế giá bất động sản. Em tin rằng việc ứng dụng mô hình hồi quy vào dự đoán giá bất động sản là một bước tiến quan trọng và hứa hẹn mang lại lợi ích lớn cho cả người mua và doanh nghiệp bất động sản, đồng thời góp phần vào sự phát triển bền vững của ngành bất động sản hiện đại.

## Nội dung phân tích

Đề này đề xuất ứng dụng hai mô hình thống kê phổ biến - mô hình hồi quy tuyến tính và mô hình hồi quy đa thức - nhằm giải quyết các bài toán định giá và dự báo giá bất động sản một cách hiệu quả. Mô hình này có vai trò quan trọng trong các hoạt động mua bán, cho thuê, thế chấp và đầu tư bất động sản.

## Mục tiêu của đề tài

Mục tiêu chính của đề án “Định giá bất động sản dựa trên mô hình hồi quy” là xác định được các yếu tố ảnh hưởng đến giá bất động sản, xây dựng mô hình phù hợp với dữ liệu đã có và cuối cùng từ đó em đánh giá độ chính xác của mô hình hồi quy và đưa ra kết luận. Mục tiêu cụ thể :

***\* Xác định các yếu tố ảnh hưởng đến giá bất động sản:***

Vị trí là một trong những yếu tố ảnh hưởng lớn nhất đến giá cả. m dựa trên những tiêu chí: Khoảng cách đến trung tâm thành phố là bao xa, khu vực cộng đồng và cơ sở hạ tầng. Tính tiện lợi trong di chuyển, giao thông quanh khu vực có thuận tiện không. Vị trí có gần các tiện ích như trường học, bệnh viện, siêu thị, công viên. Đề từ đó, em đưa ra kết luận các bất động sản có vị trí đẹp, thuận tiện đường có giá cao hơn.

Theo sau vị trí là diện tích và số phòng. Diện tích đất, diện tích sàn xây dựng ảnh hưởng hay tỷ lệ thuận với giá bất động sản hay không, hướng đất như thế nào. Miếng đất có diện tích lớn thường được ưa chuộng và có mức giá cao hơn không. Số phòng ngủ, phòng khách, phòng tắm là các yếu tố quan trọng trong định giá không. Những bất động sản có số phòng nhiều hơn thường giá có cao hơn không.

Cuối cùng em dựa trên vấn đề tiện ích hạ tầng , pháp lý và quy hoạch. Để đánh giá sự hiện diện và chất lượng của các tiện ích như công viên, sân chơi, trường học, bệnh viện ảnh hướng đến giá như thế nào. Hệ thống giao thông, hạ tầng kỹ thuật cũng là các yếu tố quan trọng. Và một phần quan trọng là tính pháp lý của bất động sản, các quy hoạch, quy định của khu vực ảnh hưởng đến giá. Bất động sản có sổ đỏ, quy hoạch rõ ràng thường được ưa chuộng có giá cao hơn không.

Việc nhận diện và phân tích ảnh hưởng của những yếu tố trên là rất quan trọng để xây dựng mô hình định giá bất động sản chính xác hơn.

***\* Xây dựng mô hình hồi quy tuyến tính và mô hình hồi quy đa thức để dự đoán giá bất động sản:***

Đánh giá độ chính xác của mô hình hồi quy thông qua các hệ số hồi quy và kiểm định của mô hình.

Việc kết hợp các phương pháp đánh giá trên sẽ cung cấp cho chúng ta một cái nhìn toàn diện về độ chính xác của mô hình hồi quy giá bất động sản. Điều này giúp xác định mức độ tin cậy của mô hình khi dự đoán giá bất động sản trong thực tế.

## Bố cục bài báo cáo

Ngoài phần mở đầu, kết luận, tài liệu tham khảo và phụ lục, nội dung của bài báo cáo còn được trình bày trong 3 chương:

*Chương 1: Tổng quan về dữ liệu lớn*

*Chương 2: Cơ sở lý thuyết và mô hình dự báo*

*Chương 3: Mô hình định giá bất động sản*

# TỔNG QUAN VỀ DỮ LIỆU LỚN

## Khái niệm dữ liệu lớn

Big Data là các tập dữ liệu có khối lượng lớn và phức tạp. Độ lớn đến mức các phần mềm xử lý dữ liệu truyền thống không có khả năng thu thập, quản lý và xử lý dữ liệu trong một khoảng thời gian hợp lý. Những tập dữ liệu lớn này có thể bao gồm các dữ liệu có cấu trúc, không có cấu trúc và bán cấu trúc, mỗi tập có thể được khai thác để tìm hiểu insights.

Dữ liệu lớn thường được mô tả qua 5V:

***Volume (Khối lượng dữ liệu):*** Dữ liệu lớn bao gồm một khối lượng dữ liệu khổng lồ, thường được đo bằng Terabyte hoặc Petabyte, vượt quá khả năng lưu trữ và xử lý của các công cụ, phương pháp truyền thống.

***Velocity (Tốc độ xử lý):*** Dữ liệu lớn được tạo ra với tốc độ nhanh chóng và liên tục từ nhiều nguồn khác nhau, đòi hỏi việc xử lý gần như thời gian thực.

***Variety (Đa dạng dữ liệu):*** Dữ liệu lớn tồn tại dưới nhiều hình thức khác nhau như cấu trúc, bán cấu trúc và phi cấu trúc, đòi hỏi các công cụ xử lý linh hoạt.

***Veracity (Tĩnh xác thực):*** Do dữ liệu lớn được thu thập từ nhiều nguồn khác nhau, chất lượng và tính chân thực của dữ liệu có thể không đồng nhất, chứa nhiều dữ liệu gây nhiễu, không nhất quán.

***Value (Giá trị dữ liệu):*** Dữ liệu lớn có tiềm năng để khai thác và phân tích giá trị ẩn có ý nghĩa kinh tế cao từ sự kết hợp của nhiều nguồn dữ liệu khác nhau.

## Một số phương pháp phân tích dữ liệu lớn

Để khai thác triệt để giá trị tiềm ẩn trong dữ liệu lớn, đã có nhiều phương pháp phân tích được phát triển và ứng dụng, trong đó phổ biến nhất:

***\* Khai phá dữ liệu (Data Mining):***

***Phân lớp (Classification):*** Dự đoán nhãn/loại của dữ liệu dựa trên các mẫu học từ dữ liệu đào tạo.

***Phân cụm (Clustering):*** Nhóm các hồ sơ dữ liệu thành các nhóm khác nhau dựa trên sự tương đồng.

***Phân tích chuỗi thời gian (Time Series Analysis):*** Xác định và phân tích các mẫu hoặc xu hướng trong chuỗi số liệu thời gian.

***Luật kết hợp (Association Rule Learning):*** Xác định quy tắc liên kết giữa các sự kiện/hành vi cùng xuất hiện trong dữ liệu.

***\* Học máy (Machine Learning):***

Học máy là một lĩnh vực của trí tuệ nhân tạo (AI), tập trung vào việc thiết kế các thuật toán và mô hình toán học cho phép máy học hỏi và dự đoán dựa trên dữ liệu đầu vào.

Các thuật toán học máy phổ biến trong phân tích dữ liệu lớn:

***Học có giám sát:*** Hồi quy tuyến tính, Hồi quy logistic, Cây quyết định, Rừng ngẫu nhiên, ...

***Học không giám sát:*** K-means, Mean-shift, Phân cụm thứ tự, Phân cụm Gaussian, ...

***Học tăng cường:*** XGBoost, LightGBM, CatBoost, …

***\* Xử lý dữ liệu luồng (Stream Processing):***

Phân tích dữ liệu liên tục theo thời gian thực đến từ các nguồn khác nhau như thiết bị IoT, mạng xã hội, mua sắm trực tuyến,…

Các công cụ xử lí dữ liệu luồng như: Apache Kafka, Apache Storm, Apache Spark Streaming,...

***\* Phân tích mạng (Network Analysis):***

Để phân tích các mối quan hệ và mẫu liên kết trực quan giữa các thực thể như người dùng, sản phẩm, website,... trong mạng lưới phức tạp.

Các kỹ thuật phân tích mạng: Trung tâm/Biên, Phát hiện cộng đồng, Xác định vai trò nút, Xác định mối quan hệ,...

***\* Trực quan hoá dữ liệu (Data Visualization):***

Biểu diễn dữ liệu lớn một cách có tổ chức, đồ họa sống động để nhận thấy các mẫu và xu hướng.

Các công cụ: Qlik, Tableau, Power BI, D3.js, ....

## Ứng dụng của phân tích dữ liệu lớn

Phân tích dữ liệu lớn đã được ứng dụng rộng rãi trong nhiều lĩnh vực nhờ khả năng khai thác thông tin hữu ích từ lượng dữ liệu khổng lồ:

***Kinh doanh và marketing:*** Phân tích hành vi khách hàng, dự đoán nhu cầu, phát hiện gian lận, cải thiện trải nghiệm khách hàng.

***Y tế và chăm sóc sức khỏe:*** Chẩn đoán bệnh chính xác, phát hiện dịch bệnh sớm, tối ưu chi phí y tế, nghiên cứu phương pháp điều trị mới.

***Giao thông vận tải:*** Tối ưu logistics, định tuyến xe, dự báo nhu cầu bảo trì hạ tầng, xây dựng hệ thống giao thông thông minh.

***Giáo dục:*** Cải thiện phương pháp giảng dạy, dự đoán kết quả học tập, nội dung giáo trình phù hợp.

***Các lĩnh vực khác:*** dự báo thời tiết, môi trường; bảo mật mạng; phòng chống tội phạm; quản lý tài nguyên hiệu quả.

Nhìn chung, phân tích dữ liệu lớn mang lại nhiều giá trị to lớn trong việc tối ưu hoạt động, nâng cao hiệu quả kinh doanh và cải thiện chất lượng cuộc sống.

## Những thách thức trong phân tích dữ liệu lớn

Mặc dù tiềm năng to lớn, phân tích dữ liệu vẫn đang đối mặt với nhiều thách thức khó khăn sau:

***Khối lượng dữ liệu khổng lồ:*** Lượng dữ liệu ngày càng lớn, dẫn đến việc lưu trữ khó khăn, đòi hỏi cơ sở hạ tầng lưu trữ cần được nâng cấp, tăng hiệu năng xử lý.

***Tốc độ dữ liệu nhanh:*** Tốc độ dữ liệu nhanh từ nhiều nguồn khiến việc đảm bảo tính nhất quán và toàn vẹn dữ liệu trở nên khó khăn. Do đó, cần hệ thống xử lý dữ liệu phân tán, song song với khả năng xử lý thời gian thực.

***Tính đa dạng của dữ liệu:*** Dữ liệu tồn tại dưới nhiều hình thức như cấu trúc, bán cấu trúc và phi cấu trúc làm phức tạp việc quản lý, lưu trữ và phân tích. Cần có cơ chế biến đổi và tích hợp dữ liệu để phân tích thống nhất.

***Bảo mật và riêng tư dữ liệu:*** Dữ liệu lớn thường chứa nhiều thông tin nhạy cảm. Việc bảo vệ quyền riêng tư và tuân thủ luật bảo mật gặp nhiều khó khăn.

***Thiếu hụt nguồn nhân lực chất lượng cao:*** Để xử lý được dữ liệu lớn cần nguồn nhân lực chất lượng cao có kiến thức chuyên môn, đồng thời kỹ năng phân tích, thống kê, học máy rất khan hiếm, khó khăn tìm kiếm và giữ chân nguồn nhân lực này.

## Các bước cơ bản trong phân tích dữ liệu lớn

Phân tích dữ liệu lớn là một quy trình bao gồm nhiều bước khác nhau. Mặc dù có thể khác nhau tùy theo bối cảnh và mục đích cụ thể, nhưng nhìn chung các bước cơ bản trong phân tích dữ liệu lớn bao gồm:

***\* Thu thập dữ liệu:***

Xác định các nguồn dữ liệu phù hợp với mục đích phân tích.

Thiết kế giải pháp để trích xuất, tích hợp dữ liệu từ nhiều nguồn khác nhau.

Đảm bảo thu thập dữ liệu đầy đủ, kịp thời và liên tục.

***\* Lưu trữ và xử lý dữ liệu:***

Xây dựng các kho dữ liệu lớn (data lake/data warehouse) phân tán để lưu trữ khối lượng dữ liệu khổng lồ.

Sử dụng hệ thống lưu trữ dữ liệu phân tán như Hadoop, NoSQL.

Tận dụng công nghệ điện toán đám mây cho khả năng mở rộng cao.

Xử lý dữ liệu song song tận dụng lợi thế của máy chủ đa nhân, GPU.

***\* Làm sạch dữ liệu:***

Loại bỏ dữ liệu nhiễu, trùng lặp, không nhất quán.

Xử lý các giá trị bị thiếu, ngoại lệ.

Chuẩn hoá và biến đổi các trường dữ liệu cho phù hợp.

Đảm bảo tính toàn vẹn và chất lượng dữ liệu trước khi phân tích.

***\* Phân tích và trích xuất thông tin:***

Lựa chọn phương pháp phân tích phù hợp: khai phá dữ liệu, học máy, xử lý dữ liệu luồng,…

Áp dụng các mô hình và thuật toán thích hợp để phân tích dữ liệu.

Trích xuất các thông tin, tri thức mới từ dữ liệu thông qua phân tích.

***\* Trình bày và diễn giải kết quả:***

Biểu diễn, trực quan hoá kết quả phân tích dưới hình thức dữ liệu, báo cáo, dashboard.

Diễn giải, giải thích ý nghĩa, bối cảnh của kết quả phân tích.

Đưa ra khuyến nghị, dự báo dựa trên kết quả phân tích.

Kết luận: Chương 1 cho chúng ta cái nhìn tổng quát nhất về dữ liệu lớn và vai trò quan trọng của nó trong thời đại số hiện nay. Qua việc phân tích chi tiết, chúng ta đã hiểu rõ hơn về các khái niệm cơ bản, phương pháp phân tích và các ứng dụng của dữ liệu lớn trong nhiều lĩnh vực khác nhau. Ngoài ra, còn tìm hiểu được về các thánh thức và các bước cơ bản trong phân tích dữ liệu lớn.

# CƠ SỞ LÝ THUYẾT VÀ MÔ HÌNH DỰ BÁO

## Tổng quan về bất động sản

### Khái niệm

***Bất động sản*** là một tài sản có giá trị bao gồm đất đai và những thứ gắn liền với đất đai, chẳng hạn như nhà cửa, cây cối, công trình xây dựng và các tài sản khác được quy định bởi pháp luật. Bất động sản đóng vai trò quan trọng trong nền kinh tế, là nguồn lực thiết yếu cho sản xuất, kinh doanh và đời sống con người.

### Phân loại bất động sản

Bất động sản được phân loại thành nhiều loại khác nhau dựa trên mục đích sử dụng và đặc điểm riêng biệt, bao gồm:

***Bất động sản nhà ở:*** Bao gồm nhà ở dân dụng, chung cư, biệt thự, nhà phố,... phục vụ cho mục đích sinh sống của con người.

***Bất động sản phi nhà ở:*** Bao gồm đất đai, nhà xưởng, kho bãi, trung tâm thương mại, văn phòng,... phục vụ cho mục đích sản xuất, kinh doanh và dịch vụ.

***Bất động sản du lịch:*** Bao gồm khách sạn, khu nghỉ dưỡng, resort,... phục vụ cho du lịch và giải trí.

***Bất động sản công nghiệp:*** Bao gồm nhà máy, khu công nghiệp,... phục vụ cho mục đích sản xuất công nghiệp.

### Đặc điểm của bất động sản

Bất động sản sở hữu những đặc điểm riêng biệt, khác biệt so với các loại tài sản khác, bao gồm:

***Tính độc đáo:*** Mỗi bất động sản đều có vị trí, diện tích, đặc điểm riêng biệt, không thể thay thế bằng bất động sản khác. Ví dụ, một căn hộ chung cư có vị trí đắc địa, tầm nhìn đẹp sẽ có giá trị cao hơn so với một căn hộ có vị trí và tầm nhìn kém.

***Tính bền vững:*** Bất động sản có tính tồn tại lâu dài, ít bị hao mòn theo thời gian. So với các loại tài sản khác như tiền mặt, vàng bạc, bất động sản có khả năng giữ giá tốt hơn và ít bị ảnh hưởng bởi biến động của thị trường.

***Tính thanh khoản:*** Bất động sản có thể chuyển đổi thành tiền mặt tương đối dễ dàng, tuy nhiên mức độ thanh khoản có thể khác nhau tùy thuộc vào loại hình bất động sản, vị trí và tình hình thị trường.

***Tính sinh lời:*** Bất động sản có thể mang lại thu nhập cho chủ sở hữu thông qua cho thuê hoặc đầu tư. Lợi nhuận từ cho thuê hoặc đầu tư bất động sản có thể mang lại nguồn thu nhập thụ động ổn định cho chủ sở hữu.

### Các yếu tố ảnh hưởng đến giá bất động sản

Giá bất động sản chịu ảnh hưởng bởi nhiều yếu tố khác nhau, bao gồm cả yếu tố tự nhiên, kinh tế, thị trường, pháp lý và yếu tố công nghệ - dịch vụ phụ trợ. Dưới đây là chi tiết về các nhóm yếu tố chính:

***\* Yếu tố tự nhiên****:*

Vị trí địa lý: Khoảng cách đến trung tâm thành phố, gần công viên, sông hồ sẽ làm tăng giá trị.

Khí hậu: Các khu vực khí hậu ôn hòa, ít thiên tai thường có giá cao hơn.

Địa hình: Đất bằng phẳng, thuận lợi xây dựng sẽ có giá cao hơn đất dốc, hiểm trở.

***\* Yếu tố kinh tế:***

Tăng trưởng kinh tế: Khi nền kinh tế phát triển, thu nhập người dân tăng sẽ kích thích nhu cầu nhà ở, làm tăng giá.

Lãi suất: Lãi suất thấp khiến vay mua nhà rẻ hơn, đẩy giá lên. Ngược lại, lãi suất cao khiến giá nhà giảm,....

***\* Yếu tố thị trường****:*

Cung cầu: Nếu cung nhà ở thấp hơn cầu sẽ đẩy giá lên cao. Ngược lại thì giá sẽ giảm.

Xu hướng dân số: Dân số tăng, ly hôn tăng sẽ làm tăng nhu cầu và giá nhà.

***\* Yếu tố pháp lý****:*

Quy hoạch, đô thị hóa: Các quy định liên quan đến sử dụng đất, cấp phép xây dựng ảnh hưởng lớn đến giá.

Luật thuế đất: Chính sách thuế đất cao hay thấp sẽ tác động trực tiếp đến giá nhà đất.

***\* Công nghệ - Dịch vụ:***

Giá bất động sản chịu ảnh hưởng từ nhiều yếu tố khác nhau, bao gồm các yếu tố kinh tế vĩ mô, tự nhiên, thị trường, pháp lý và cả yếu tố công nghệ-dịch vụ của khu vực đó. Sự tác động của các yếu tố này thay đổi theo không gian và thời gian, làm cho việc dự đoán giá bất động sản trở nên phức tạp.

## Mô hình dự báo

### Hồi quy tuyến tính:

#### Khái niệm:

Hồi quy tuyến tính là một phương pháp thống kê được sử dụng để mô tả mối quan hệ tuyến tính giữa một biến phụ thuộc (y) và một hoặc nhiều biến độc lập (x). Mục tiêu của hồi quy tuyến tính là tìm ra một hàm tuyến tính f(x) để dự đoán giá trị của y dựa trên giá trị của x.

#### Công thức:

Hàm tuyến tính f(x) trong hồi quy tuyến tính được biểu thị bởi công thức sau:

f(x) = β0 + β1x1 + β2x2 + ... + βkxk

Trong đó:

β0, β1, β2, ..., βk là các hệ số hồi quy cần được ước tính.

x1, x2, ..., xk là các biến độc lập.

#### Ước tính hệ số hồi quy:

Có nhiều phương pháp khác nhau để ước tính hệ số hồi quy trong hồi quy tuyến tính. Phương pháp phổ biến nhất là phương pháp bình phương tối thiểu (Ordinary Least Squares - OLS). Phương pháp OLS tìm ra các giá trị của β0, β1, β2, ..., βk sao cho tổng bình phương sai số giữa giá trị dự đoán f(x) và giá trị thực tế y được thu nhỏ nhất.

### Hồi quy đa thức

#### Khái niệm

Hồi quy đa thức là một phương pháp thống kê được sử dụng để mô tả mối quan hệ giữa một biến phụ thuộc (Y) và một hoặc nhiều biến độc lập (X). Trong hồi quy đa thức, mối quan hệ giữa các biến được mô tả bởi một hàm đa thức bậc p, trong đó p là bậc của đa thức.

#### Công thức:

Mô hình hồi quy đa thức được biểu diễn bằng công thức sau:

Y = β0 + β1X + β2X^2 + ... + βpX^p + ε

Trong đó:

Y: Biến phụ thuộc

X: Biến độc lập

β0, β1, β2, ..., βp: Hệ số hồi quy

ε: Sai số ngẫu nhiên

#### Ước lượng hệ số hồi quy:

Có nhiều phương pháp khác nhau để ước lượng hệ số hồi quy trong mô hình hồi quy đa thức. Một trong những phương pháp phổ biến nhất là phương pháp bình phương nhỏ nhất tuyến tính (OLS). Phương pháp OLS tìm kiếm các giá trị của hệ số hồi quy β0, β1, β2, ..., βp sao cho tổng bình phương sai số (RSS) được thu nhỏ nhất.

### Đánh giá mô hình:

Sau khi ước lượng được hệ số hồi quy, cần đánh giá xem mô hình có phù hợp với dữ liệu hay không. Có nhiều chỉ số thống kê khác nhau được sử dụng để đánh giá mô hình, bao gồm:

***Hệ số xác định ():*** đo lường mức độ mà mô hình hồi quy giải thích được sự biến thiên của dữ liệu đầu vào, cho biết tỷ lệ phần trăm biến đổi của biến phụ thuộc được giải thích bởi các biến độc lập.

Trong đó: yi là giá trị thực tế, là giá trị dự đoán, là giá trị trung bình của các giá trị thực tế, n là số lượng mẫu

***Mean Absolute Error (MAE):*** là chỉ số đo lường mức độ sai số trung bình của các dự đoán so với các giá trị thực tế. Nó được tính bằng cách lấy giá trị tuyệt đối của sai số (tức là khoảng cách giữa giá trị dự đoán và giá trị thực tế) và sau đó tính trung bình.

MAE =

***Mean Squared Error (MSE):*** là chỉ số đo lường mức độ sai số trung bình của các dự đoán so với các giá trị thực tế bằng cách lấy bình phương của sai số và sau đó tính trung bình.

***Root Mean Squared Error (RMSE)*** là căn bậc hai của MSE, giúp giảm đơn vị của sai số về cùng thang đo với giá trị dự đoán và giá trị thực tế.

Các chỉ số MAE, MSE và RMSE đều cho thấy mức độ sai số trung bình của mô hình. Các chỉ số này càng nhỏ, mô hình càng chính xác.

## Tóm tắt các nghiên cứu trước

**Một số nghiên cứu tiêu biểu**

***Palm et al. (2003):*** Sử dụng hồi quy đa biến để định giá bất động sản ở khu vực Chicago, Hoa Kỳ. Kết quả nghiên cứu cho thấy mô hình hồi quy đa biến có độ chính xác cao hơn so với mô hình hồi quy tuyến tính.

***Miller & Wang (2006):*** Sử dụng mạng nơ-ron nhân tạo để định giá bất động sản ở khu vực Boston, Hoa Kỳ. Kết quả nghiên cứu cho thấy mạng nơ-ron nhân tạo có thể dự đoán giá bất động sản chính xác hơn so với các phương pháp truyền thống.

***He et al. (2017):*** Sử dụng dữ liệu lớn và GIS để định giá bất động sản ở khu vực Bắc Kinh, Trung Quốc. Kết quả nghiên cứu cho thấy mô hình sử dụng dữ liệu lớn và GIS có độ chính xác cao hơn so với các mô hình truyền thống.

Kết luận: Chương 2 đã cung cấp cái nhìn tổng quan về bất động sản và hai mô hình dự báo hồi quy tuyến tính và hồi quy đa thức. Bằng việc tổng hợp các lý thuyết và thực tiễn về định giá bất động sản, chương này đã đặt nền tảng vững chắc cho các phân tích chuyên sâu hơn ở các chương sau.

# MÔ HÌNH ĐỊNH GIÁ BẤT ĐỘNG SẢN

## Giới thiệu dữ liệu

### Tập dữ liệu

Thu thập dữ liệu là bước nền tảng và quan trọng trong quá trình xây dựng mô hình hồi quy để định giá bất động sản. Trong bài phân tích này tôi sử dụng bộ dữ liệu lịch sử thị trường về định giá bất động sản được thu thập từ Quận Sindian, thành phố Tân Bắc, Đài Loan. Bộ dữ liệu được đóng góp ngày 17/08/2018 trên trang web UCI Machine Learning Repository bao gồm các thông tin sau:

X1: Transaction date

X2: House age

X3: Distance to the nearest MRT station

X4: Number of convenience stores

X5: Latitude

X6: Longitude

Y: House price of unit area

Sau khi có được bộ dữ liệu tôi xác định “Y: House price of unit area” là biến phụ thuộc, các biến “X1: Transaction date”, “X2: House age”, “X3: Distance to the nearest MRT station”, “X4: Number of convenience stores”, “X5: Latitude”, “X6: Longitude” là các biến độc lập. Việc xác định và hiểu rõ biến phụ thuộc và biến độc lập là bước đầu tiên và quan trọng trong quá trình xây dựng mô hình hồi quy. Biến phụ thuộc “Y” là mục tiêu mà tôi muốn dự đoán, trong khi các biến độc lập X1, X2, X3, X4, X5, X6 là những yếu tố mà tôi tin rằng có thể ảnh hưởng đến biến phụ thuộc. Qua đó, mô hình hồi quy có thể giúp tôi hiểu rõ hơn về mối quan hệ giữa các yếu tố này và đưa ra những dự đoán chính xác.

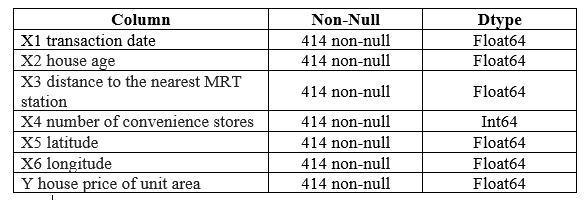
### Làm sạch dữ liệu

Để đảm bảo rằng dữ liệu sử dụng cho mô hình hồi quy là sạch, nhất quán và không có lỗi. Tôi tiến hành cách bước làm sạch dữ liệu:

#### Bước 1: Xử lý dữ liệu thiếu:

Trong bộ dữ liệu, tôi sử dụng câu lệnh “df.infor()” để biết định dạng và số lượng quan sát not-null của mỗi trường trong dataframe.

Bảng 3.1. Xử lý dữ liệu thiếu



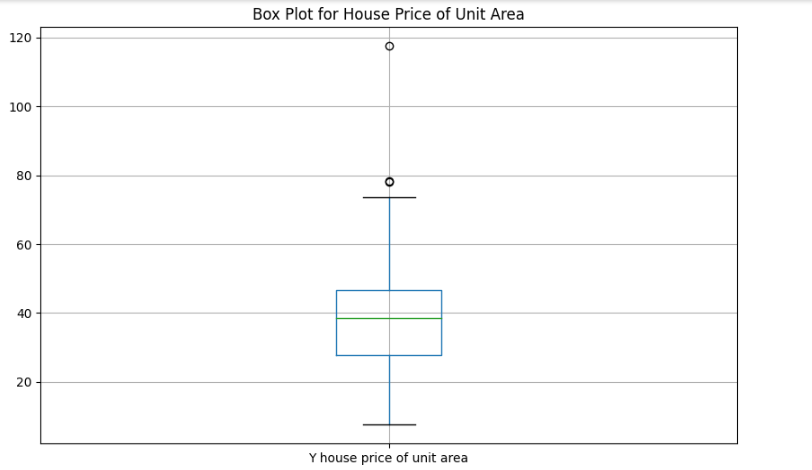
Quá trình kiểm tra bộ dữ liệu, tôi nhận thấy rằng không có giá trị nào bị thiếu. Tất cả các biến đều đã được điền đầy đủ và không cần thực hiện các bước loại bỏ giá trị thiếu và điền giá trị thiếu. Do đó, tôi có thể chuyển sang bước tiếp theo trong quá trình làm sạch dữ liệu.

#### Bước 2: Xử lý giá trị ngoại lai

***\* Xác định các giá trị ngoại lai (outliers):***

***Phân tích hộp (Box Plot):*** Sử dụng biểu đồ hộp để trực quan hóa phân phối dữ liệu và xác định các giá trị ngoại lai. Các điểm nằm ngoài phạm vi [Q1 - 1.5 \* IQR, Q3 + 1.5 \* IQR] (Q1: phần tư thứ nhất, Q3: phần tư thứ ba, IQR: khoảng tứ phân vị) thường được coi là giá trị ngoại lai.

Đối với bộ dữ liệu, tôi tiến hành sử dụng Box Plot để xác định các hàng ngoại lai. Kết quả thu được 3 hàng 220, 270 và 312 có giá trị ngoại lai lần lượt là 78.3, 117.5, 78.0



Hình 3.1. Biểu đồ Box Plot để xác định giá trị ngoại lai

***Xử lý các giá trị ngoại lai:*** Các giá trị ngoại lai có thể được xử lý bằng cách: Loại bỏ các giá trị ngoại lai, Biến đổi dữ liệu và Gán giá trị mới

Đối với 3 giá trị ngoại lại trong bộ dữ liệu, tôi thay thế bằng giá trị trung bình của cột ‘Y house price of unit area’

### Biến đổi dữ liệu

#### Chuẩn hóa dữ liệu (normalization)

Các biến trong dữ liệu có đơn vị và khoảng giá trị khác nhau. Ví dụ, "X3 distance to the nearest MRT station" nằm trong khoảng từ vài chục đến vài trăm nghìn kilomet, trong khi "Y house price of unit area" nằm trong khoảng từ vài chục USD.

Khi các biến có phạm vi giá trị rất khác nhau, các mô hình có thể trở nên nhạy cảm với các biến có giá trị lớn hơn và không thể tối ưu hóa đúng cách. Vì thế tôi chuẩn hóa các biến “X1”, “X2” , “X3”, “X4” và “Y” bằng Min-Max Scaling.

Chuyển đổi các giá trị của biến về khoảng [0, 1].

Công thức: X′ = (X−Xmin) / (Xmax−Xmin)

**Trong đó:**

**X:** Giá trị gốc của biến.

**X′:** Giá trị đã được chuẩn hóa.

**X(min):** Giá trị nhỏ nhất của biến trong tập dữ liệu.

**X(max):** Giá trị lớn nhất của biến trong tập dữ liệu.

#### Biến đổi dữ liệu không tuyến tính

Biến đổi dữ liệu không tuyến tính là để làm cho mối quan hệ giữa các biến đầu vào và biến mục tiêu trở nên tuyến tính hơn, cải thiện hiệu suất của các mô hình tuyến tính và xử lý các vấn đề liên quan đến phân phối của dữ liệu.

Tôi sử dụng phương pháp “Log Transformation” để biến đổi dữ liệu không tuyến tính cột ‘Y house price of unit area’

Dùng để xử lý dữ liệu có phân phối lệch phải, biến đổi các giá trị lớn thành các giá trị nhỏ hơn.

Công thức: X′=log(X+1)

Sử dụng log(X+1) để tránh log của giá trị 0.

### Kiểm tra và loại bỏ đa cộng tuyến (multicollinearity)

Kiểm tra và loại bỏ đa cộng tuyến là một bước quan trọng trong quy trình tiền xử lý dữ liệu, giúp đảm bảo tính ổn định và độ tin cậy của mô hình hồi quy. Các phương pháp phổ biến bao gồm kiểm tra hệ số phóng đại phương sai (VIF) và ma trận tương quan, và các phương pháp loại bỏ đa cộng tuyến như loại bỏ biến số, kết hợp các biến số, và sử dụng phân tích thành phần chính (PCA).

Trong báo cáo tôi sử dụng hệ số phóng đại phương sai (Variance Inflation Factor - VIF) để kiểm tra đa cộng tuyến. VIF đo lường mức độ mà một biến số có thể được dự đoán từ các biến số khác.

Công thức tính VIF cho biến *i*: VIFi=1/(1−Ri^2)

Trong đó Ri^2 là hệ số xác định của hồi quy biến *i* trên các biến độc lập khác.

Quy tắc thông thường:

VIF < 5: Đa cộng tuyến không đáng kể.

5 ≤ VIF < 10: Đa cộng tuyến trung bình.

VIF ≥ 10: Đa cộng tuyến cao.

Bảng hiển thị VIF cho từng biến số

Bảng 3.2. Bảng hiển thị VIF cho từng biến số

|  |  |
| --- | --- |
| **VIF của các biến số** | **VIF** |
| X1 transaction date | 3.982447 |
| X2 house age | 3.470895 |
| X3 distance to the nearest MRT station | 2.946550 |
| X4 number of convenience stores | 4.742453 |
| X5 latitude | 5.946057 |
| X6 longitude | 5.945190 |

Kết quả VIF không có bất kỳ biến số nào cao (lớn hơn 10) như vậy kết luận được không có đa cộng tuyến trong mô hình.

## Khám phá dữ liệu (Exploratory Data Analysis - EDA)

Ma trận tương quan là một công cụ thống kê giúp ta hiểu được mối quan hệ tuyến tính giữa các biến số.

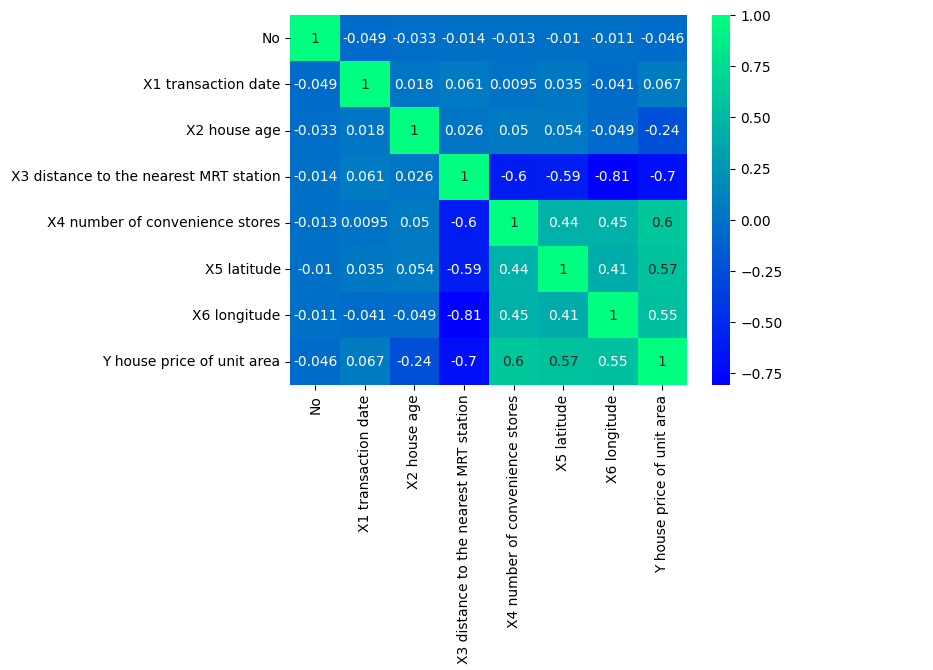
Kết quả của ma trận tương quan sẽ cho biết mức độ liên quan giữa các biến số trong dữ liệu. Các giá trị tương quan nằm trong khoảng từ -1 đến 1:

1: Mối quan hệ tương quan hoàn hảo dương

-1: Mối quan hệ tương quan hoàn hảo âm

0: Không có mối quan hệ tương quan

### Ma trận Tương quan với Giá Nhà



Hình 3.2 Ma trận Tương quan với Giá nhà

***X3 (distance to the nearest MRT station)****:* Mối tương quan âm mạnh với giá nhà (-0.67). Khoảng cách đến trạm MRT càng xa, giá nhà càng thấp.

***X4 (number of convenience stores)****:* Mối tương quan dương mạnh với giá nhà (0.57). Số lượng cửa hàng tiện lợi càng nhiều, giá nhà càng cao.

***X5 (latitude) và X6 (longitude):*** Mối tương quan dương với giá nhà (0.55 và 0.52). Vị trí địa lý (vĩ độ và kinh độ) có ảnh hưởng tích cực đến giá nhà.

***X2 (house age):*** Mối tương quan âm nhẹ với giá nhà (-0.21). Nhà càng cũ, giá nhà càng giảm.

### Mối Quan Hệ Giữa Các Biến Số Khác

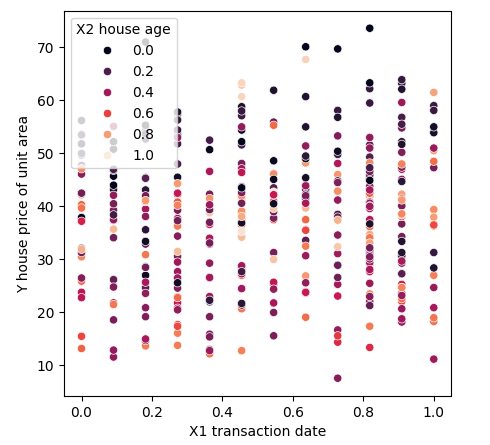
***X3 (distance to the nearest MRT station)****:*Tương quan âm rất mạnh với X6 (longitude) (-0.81) và X4 (number of convenience stores) (-0.60). Các trạm MRT thường ở khu vực phát triển, gần nhiều cửa hàng tiện lợi và ở vị trí địa lý cụ thể.

***X5 (latitude)*** *và* ***X4 (number of convenience stores)****:* Mối tương quan dương (0.44). Khu vực có nhiều cửa hàng tiện lợi thường nằm ở một vĩ độ cụ thể.

Biểu đồ phân tán (scatter plots) để xem xét mối quan hệ giữa các biến số và giá.

Biểu đồ scatter: Hiển thị mối quan hệ giữa giá nhà (Y) với các biến độc lập như tuổi nhà (X2), khoảng cách đến ga tàu điện ngầm (X3), số cửa hàng tiện lợi (X4), vĩ độ (X5), kinh độ (X6).

#### Mối quan hệ giữa giá nhà và ngày giao dịch:



Hình 3.3 Biểu đồ thể hiện mối quan hệ giữa giá nhà và ngày giao dịch

Biểu đồ trên thể hiện mối quan hệ giữa giá nhà trên mỗi đơn vị diện tích (Y house price of unit area) và ngày giao dịch (X1 transaction date), với màu sắc của điểm biểu diễn tuổi của ngôi nhà (X2 house age). Dưới đây là một số nhận xét về biểu đồ:

Biểu đồ không cho thấy một xu hướng rõ ràng giữa giá nhà và ngày giao dịch. Các điểm dữ liệu phân bố khá đều và không có sự tăng giảm rõ rệt theo ngày giao dịch. Giá nhà có sự biến động lớn trong suốt khoảng thời gian giao dịch từ 0.0 đến 1.0.

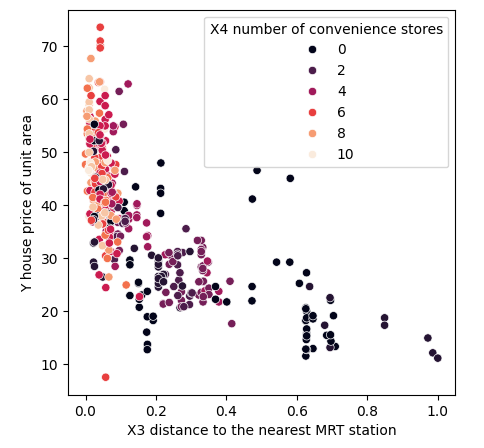
***\* Phân bố giá nhà:***

Giá nhà trên mỗi đơn vị diện tích phân bố từ khoảng 10 đến hơn 70. Các mức giá nhà cao (trên 50) được ghi nhận ở mọi khoảng thời gian giao dịch, nhưng không có một xu hướng rõ ràng nào về sự thay đổi giá nhà theo thời gian.

***\* Tuổi của ngôi nhà:***

Các điểm dữ liệu được mã hóa màu sắc biểu diễn tuổi của ngôi nhà từ 0.0 đến 1.0. Các ngôi nhà có tuổi khác nhau dường như phân bố đồng đều trên toàn bộ khoảng giá và ngày giao dịch. Không có sự tập trung rõ rệt của ngôi nhà có tuổi cụ thể nào tại một khoảng giá hay ngày giao dịch nhất định.

#### Mối quan hệ giữa giá nhà và khoảng cách tới ga MRT:



Hình 3.4 Biểu đồ thể hiện mối quan hệ giữa giá nhà và khoảng cách tới ga MRT

Biểu đồ trên thể hiện mối quan hệ giữa giá nhà trên mỗi đơn vị diện tích (Y house price of unit area) và khoảng cách tới ga MRT gần nhất (X3 distance to the nearest MRT station), với màu sắc của điểm biểu diễn số lượng cửa hàng tiện lợi trong khu vực (X4 number of convenience stores). Dưới đây là một số nhận xét về biểu đồ:

Biểu đồ cho thấy xu hướng giá nhà giảm khi khoảng cách tới ga MRT tăng lên. Các điểm dữ liệu tập trung nhiều ở khoảng cách gần với ga MRT (dưới 0.2) và giá nhà cao hơn. Ở khoảng cách lớn hơn (trên 0.4), giá nhà có xu hướng thấp hơn, và các điểm dữ liệu trở nên thưa thớt hơn.

***\* Phân bố giá nhà:***

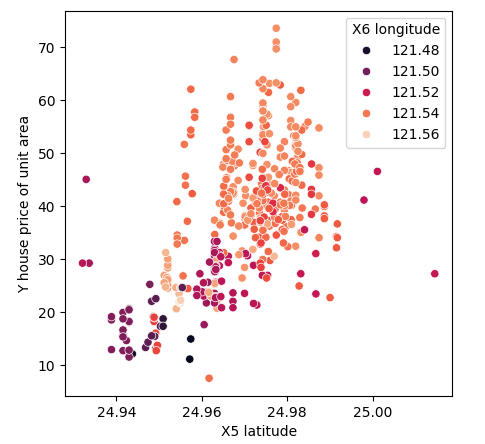
Giá nhà trên mỗi đơn vị diện tích tập trung nhiều trong khoảng từ 20 đến 60 đối với các khoảng cách gần ga MRT. Ở khoảng cách xa hơn, giá nhà chủ yếu nằm trong khoảng từ 10 đến 40.

***\* Số lượng cửa hàng tiện lợi:***

Các điểm dữ liệu có nhiều màu sắc khác nhau, biểu diễn số lượng cửa hàng tiện lợi khác nhau. Các khu vực có nhiều cửa hàng tiện lợi hơn (màu sắc từ đỏ đến cam) dường như tập trung ở khoảng cách gần ga MRT và có giá nhà cao hơn. Các điểm dữ liệu có ít cửa hàng tiện lợi (màu đen và tím) có xu hướng phân bố rộng hơn và có giá nhà thấp hơn ở khoảng cách xa ga MRT.

#### Mối quan hệ giữa giá nhà và vĩ độ:

Hình 3.5 Biểu đồ mối quan hệ giữa giá nhà và vĩ độ



Biểu đồ trên thể hiện mối quan hệ giữa giá nhà trên mỗi đơn vị diện tích (Y house price of unit area) và vĩ độ (X5 latitude), với màu sắc của điểm biểu diễn kinh độ (X6 longitude). Dưới đây là một số nhận xét về biểu đồ:

Biểu đồ cho thấy xu hướng giá nhà tăng khi vĩ độ tăng. Các điểm dữ liệu tập trung nhiều ở vĩ độ từ 24.96 đến 25.00 và giá nhà cao hơn. Giá nhà dưới 30 chủ yếu tập trung ở vĩ độ thấp hơn (dưới 24.96).

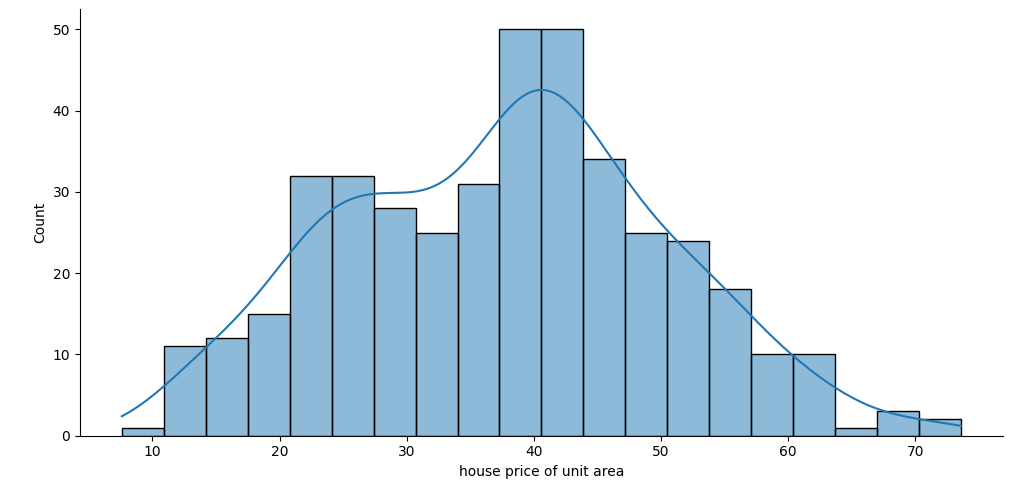
***\* Phân bố giá nhà:***

Giá nhà trên mỗi đơn vị diện tích phân bố trong khoảng từ 10 đến 70, với phần lớn các giá nhà tập trung trong khoảng từ 20 đến 50. Các mức giá nhà cao (trên 50) xuất hiện nhiều ở vĩ độ từ 24.96 đến 25.00.

***\* Kinh độ:***

Các điểm dữ liệu được mã hóa màu sắc biểu diễn kinh độ. Nhìn chung, các điểm có màu sắc khác nhau phân bố đều trên toàn bộ biểu đồ. Các điểm dữ liệu có kinh độ cao hơn (màu từ cam đến đỏ nhạt) có xu hướng tập trung ở vĩ độ cao hơn (gần 25.00), và giá nhà ở đây cũng có xu hướng cao hơn. Các điểm dữ liệu có kinh độ thấp hơn (màu từ đen đến tím) có xu hướng phân bố ở vĩ độ thấp hơn (dưới 24.96), và giá nhà ở đây phần lớn thấp hơn.

#### Phân bố giá nhà trên mỗi đơn vị diện tích



Hình 3.6 Biểu đồ phân bố giá nhà trên mỗi đơn vị diện tích

Biểu đồ trên là biểu đồ histogram kết hợp với đường cong mật độ (density curve) thể hiện phân bố giá nhà trên mỗi đơn vị diện tích. Dưới đây là một số nhận xét về biểu đồ:

***\* Hình dạng phân phối:***

Biểu đồ cho thấy phân phối của giá nhà có dạng gần đối xứng với một đỉnh (unimodal). Đỉnh của phân phối nằm ở khoảng giá trị từ 35 đến 45, nghĩa là phần lớn các giá nhà trên mỗi đơn vị diện tích tập trung ở mức này.

***\* Tần suất:***

Khoảng giá trị từ 35 đến 45 có tần suất cao nhất với số lượng khoảng 50. Các khoảng giá trị từ 10 đến 70 có sự phân bố rộng rãi với tần suất giảm dần ở hai đầu.

***\* Phân tán:***

Phân phối có độ phân tán khá lớn, trải dài từ khoảng 10 đến khoảng 70. Điều này cho thấy sự biến động lớn về giá nhà trên mỗi đơn vị diện tích trong dữ liệu.

***\* Đường cong mật độ:***

Đường cong mật độ cho thấy phân phối gần như đối xứng với đỉnh cao nhất tại khoảng 40. Đường cong mật độ mịn, giúp xác định rõ ràng sự biến thiên của giá nhà trên mỗi đơn vị diện tích. Đường KDE cung cấp một ước lượng trơn tru về mật độ xác suất của dữ liệu. Đỉnh của đường KDE phù hợp với mode của biểu đồ histogram, củng cố rằng phần lớn các điểm dữ liệu rơi vào giá trị này.

## Xây dựng mô hình hồi quy

Mô hình hồi quy tuyến tính ban đầu được sử dụng để dự đoán giá nhà(Y) là:

**Y = b0 + b1X1 + b2X2 + b3X3 + b4X4 + b5X5 + b6X6**

Mô hình hồi quy đa thức ban đầu được sử dụng để dự đoán giá nhà(Y) là:

**Y = β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5 + β6X6 + β7X1^2… + ε**

Tôi chia bộ dữ liệu thành tập Train và tập Test

Trong đó:

Tập Train: Y

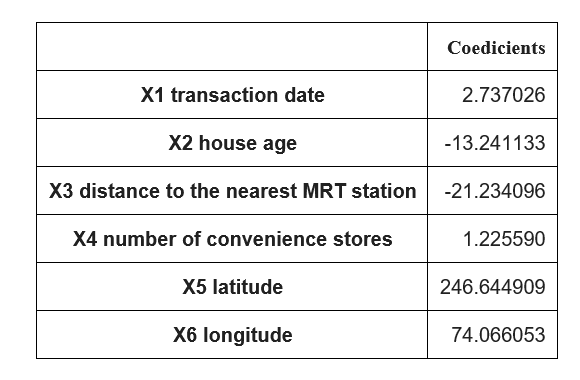
Tập Test: X1, X2, X3, X4, X5, X6

**Cách hoạt động**: Hàm **train\_test\_split** chia ngẫu nhiên dữ liệu thành hai tập với tỷ lệ được chỉ định bởi **test\_size**. Tập huấn luyện (70% dữ liệu) sẽ được sử dụng để huấn luyện mô hình hồi quy, trong khi tập kiểm tra (30% dữ liệu) sẽ được sử dụng để đánh giá hiệu suất của mô hình đã được huấn luyện.

### Mô hình hồi quy tuyến tính

Dựa vào hệ số của mô hình hồi quy tuyến tính, ta có thể đánh giá sự ảnh hưởng của từng biến đến giá nhà trên mỗi đơn vị diện tích.

Bảng 3.3. Các hệ số của mô hình



***\* Đánh giá các hệ số:***

***X1 (transaction date)****:* Hệ số dương (2.737026) cho thấy rằng khi thời gian giao dịch tăng, giá nhà có xu hướng tăng.

***X2 (house age)****:* Hệ số âm (-13.241133) biểu thị rằng khi tuổi của ngôi nhà tăng, giá nhà có xu hướng giảm.

***X3 (distance to the nearest MRT station)****:* Hệ số âm (-21.234096) cho thấy rằng khi khoảng cách đến trạm MRT gần nhất tăng, giá nhà có xu hướng giảm. Điều này hợp lý vì gần các phương tiện giao thông công cộng thường có giá trị cao hơn.

***X4 (number of convenience stores)****:* Hệ số dương (1.225590) cho thấy rằng khi số lượng cửa hàng tiện lợi tăng, giá nhà cũng có xu hướng tăng.

***X5 (latitude)****:* Hệ số dương lớn (246.644909) cho thấy rằng khi vĩ độ tăng, giá nhà có xu hướng tăng mạnh. Điều này có thể phản ánh vị trí địa lý và sự phát triển khu vực.

***X6 (longitude)****:* Hệ số dương (74.066053) cho thấy rằng khi kinh độ tăng, giá nhà cũng có xu hướng tăng.

Từ các hệ số trên, chúng ta thấy rằng:

***Tăng giá:*** Thời gian giao dịch, số lượng cửa hàng tiện lợi, vĩ độ, và kinh độ có tác động tích cực đến giá nhà trên mỗi đơn vị diện tích.

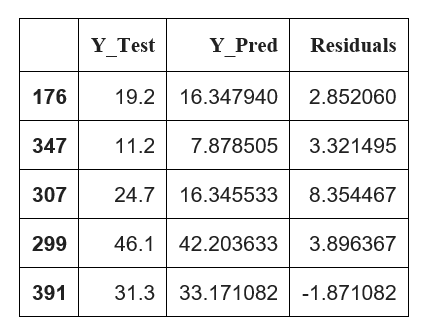
***Giảm giá:*** Tuổi của ngôi nhà và khoảng cách đến trạm MRT gần nhất có tác động tiêu cực đến giá nhà trên mỗi đơn vị diện tích.

Mô hình hồi quy tuyến tính này cho thấy rằng các yếu tố như thời gian giao dịch, số lượng cửa hàng tiện lợi, vĩ độ và kinh độ có tác động tích cực đến giá nhà. Ngược lại, tuổi của ngôi nhà và khoảng cách đến trạm MRT gần nhất có tác động tiêu cực đến giá nhà. Việc hiểu rõ các yếu tố này sẽ giúp trong việc định giá và đầu tư bất động sản.

### Mô hình hồi quy đa thức:

Dựa trên bảng kết quả bao gồm giá trị thực tế (Y\_Test), giá trị dự đoán (Y\_Pred) và residuals (phần dư), chúng ta có thể đánh giá chi tiết hơn về hiệu suất của mô hình hồi quy đa thức.

Bảng 3.4. Mô hình hồi quy đa thức



Giá trị sai số (Residuals) cho thấy mức độ chênh lệch giữa giá trị dự đoán (Y\_Pred) và giá trị thực tế (Y\_Test).

***\* Sai số dự đoán****:*

Sai số dự đoán là hiệu số giữa giá trị thực tế và giá trị dự đoán. Các giá trị residuals cho thấy có sự chênh lệch giữa giá trị thực tế và giá trị dự đoán, dao động từ -1.871082 đến 8.354467.

Giá trị residuals dương (2.852060, 3.321495, 8.354467, 3.896367) cho thấy mô hình dự đoán thấp hơn giá trị thực tế.

Giá trị residuals âm (-1.871082) cho thấy mô hình dự đoán cao hơn giá trị thực tế.

Mô hình hồi quy đa thức có thể dự đoán giá nhà với một mức độ chính xác nhất định, nhưng vẫn cần cải thiện để giảm thiểu sai số dự đoán, đặc biệt trong những trường hợp có sai số lớn.

## Phân tích kết quả hồi quy

### Phân tích kết quả mô hình hồi quy tuyến tính

#### Đánh Giá Tổng Quan Hiệu Suất Mô Hình

Giá trị hệ số R² (Coefficient of Determination) của mô hình là 0.649, điều này có nghĩa là mô hình hồi quy của tôi giải thích được 64.9% sự biến động của giá nhà trên đơn vị diện tích. Đây là một kết quả tương đối tốt, cho thấy mô hình có khả năng dự đoán khá chính xác. Tuy nhiên, còn khoảng 35.1% sự biến động chưa được giải thích, có thể là do các yếu tố khác chưa được mô hình hóa hoặc các yếu tố ngẫu nhiên khác.

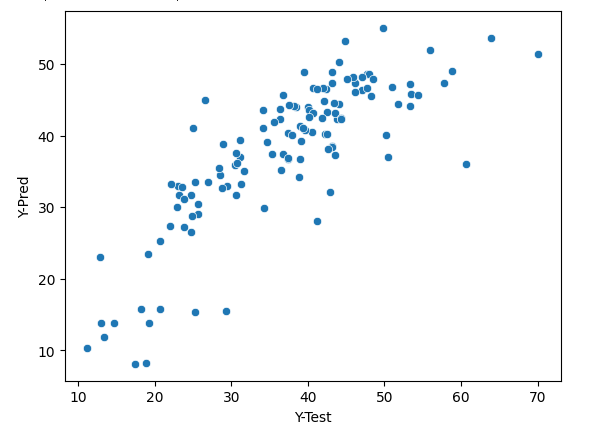
Giá trị MAE (Mean Absolute Error) là 5.5334487, điều này có nghĩa là trung bình các dự đoán của mô hình sai lệch khoảng 5.334487 đơn vị so với giá trị thực tế. Giá trị MAE càng thấp, hiệu suất của mô hình càng cao. Giá trị này nhỏ cho thấy rằng các dự đoán không bị ảnh hưởng nhiều bởi các giá trị ngoại lệ hoặc các sai số lớn.

Giá trị MSE (Root Mean Squared Error) là 46.460281, điều này có nghĩa là trung bình bình phương của sai số giữa giá trị dự đoán và giá trị thực tế là 46.460281 đơn vị. Một giá trị MSE thấp chỉ ra rằng mô hình có hiệu suất tốt. Vì sai số được bình phương, nên MSE nhạy cảm hơn với các giá trị ngoại lệ so với MAE. Giá trị này nhỏ cho thấy rằng các dự đoán không bị ảnh hưởng nhiều bởi các giá trị ngoại lệ hoặc các sai số lớn.

Giá trị RMSE (Root Mean Squared Error) là 6.816178 có nghĩa là trung bình sai số dự đoán của mô hình là khoảng 6.816178 đơn vị so với giá trị thực tế. RMSE nhạy cảm với các ngoại lệ và cho thấy mức độ phân tán của các sai số. Giá trị RMSE thấp cho thấy mô hình có khả năng dự đoán tốt và ít sai số.

Các chỉ số này đều cho thấy rằng mô hình hồi quy của tôi có khả năng dự đoán tốt giá trị mục tiêu với độ chính xác cao. Giá trị R² cho thấy mô hình giải thích được phần lớn sự biến động của dữ liệu. Các giá trị MAE, MSE và RMSE đều nhỏ, cho thấy sai số dự đoán của mô hình là khá thấp, đồng thời không có sai số lớn nào chi phối.

#### Phân Tích Biểu Đồ Phân Tán Giá Trị Thực Tế và Giá Trị Dự Đoán



Hình 3.7 Biểu đồ Phân Tán Giữa Giá Trị Thực Tế và Giá Trị Dự Đoán

Biểu đồ phân tán này cho thấy sự phân bố của các giá trị dự đoán so với giá trị thực tế. Nếu mô hình hồi quy có hiệu suất tốt, các điểm dữ liệu sẽ nằm gần đường thẳng có độ dốc là 1 (đường y = x), cho thấy rằng giá trị dự đoán gần giống với giá trị thực tế.

***Trục X (Y-Test):*** Biểu thị giá trị thực tế của biến mục tiêu trong tập dữ liệu kiểm tra.

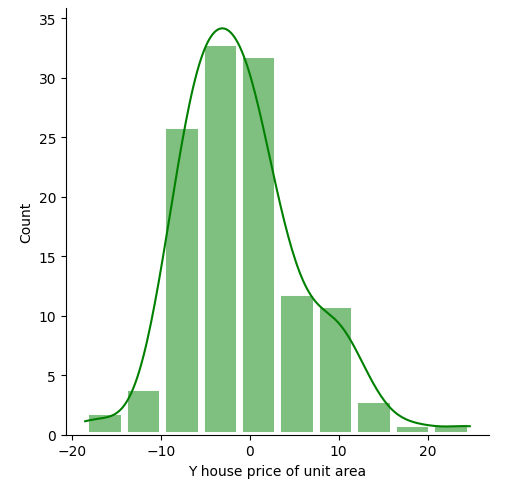
***Trục Y (Y-Pred):*** Biểu thị giá trị dự đoán của biến mục tiêu bởi mô hình hồi quy

Trong biểu đồ này, Phần lớn các điểm dữ liệu nằm gần đường chéo (đường y = x), biểu thị rằng giá nhà dự đoán khá khớp với giá nhà thực tế. Tuy nhiên, có Giá nhà thực tế và dự đoán trải dài từ khoảng 10 đến 70.

Các giá nhà cao hơn (trên 50) cũng được dự đoán khá chính xác, nhưng có một số ít điểm bị dự đoán thấp hơn thực tế. Phần lớn các điểm dữ liệu tập trung ở khoảng giá từ 20 đến 50, cho thấy đây là khoảng giá phổ biến nhất trong dữ liệu.

Mặc dù có một số điểm nằm rời rạc, nhưng nhìn chung mô hình dự đoán hoạt động tốt ở hầu hết các khoảng giá.một số điểm dữ liệu nằm xa đường chéo, cho thấy có những dự đoán không chính xác. Điều này có thể do các yếu tố không được mô hình hóa hoặc các biến động bất thường trong dữ liệu.

#### Phân tích phân phối phần dư



Hình 3.8 Biểu đồ phân phối phần dư

Biểu đồ này hiển thị số lượng phần dư nằm trong mỗi khoảng giá trị và đường cong KDE mô tả phân phối xác suất của các phần dư.

***Trục X (Y house price of unit area):*** Biểu thị giá trị của phần dư.

***Trục Y (Count):*** Biểu thị số lượng các phần dư trong mỗi khoảng giá trị cụ thể.

Biểu đồ phân phối phần dư cho thấy phần lớn các phần dư tập trung quanh giá trị 0, tạo thành phân phối gần giống phân phối chuẩn. Điều này chỉ ra rằng mô hình không có xu hướng hệ thống trong việc dự đoán cao hơn hoặc thấp hơn so với giá trị thực tế. Hình dạng của biểu đồ histogram gần giống với hình chuông, nghĩa là phần dư có xu hướng phân phối theo phân phối chuẩn (normal distribution). Đây là một dấu hiệu tích cực, các giả định về phân phối phần dư của mô hình hồi quy tuyến tính có thể được đáp ứng.

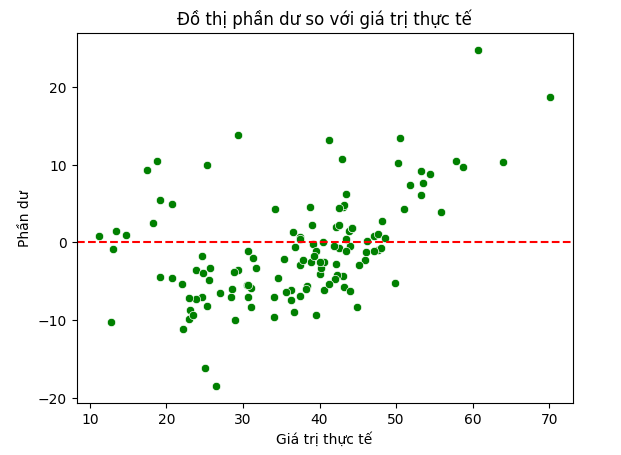
Khoảng giá trị từ -5 đến 5 có tần suất cao nhất, khoảng 30 đến 35 lần, cho thấy phần lớn các sai số dự đoán nằm trong khoảng này. Sai số dự đoán từ -15 đến -10 và 10 đến 15 có tần suất thấp hơn, khoảng dưới 5 lần, cho thấy có ít dự đoán sai lệch lớn.

Phân phối có độ phân tán khá lớn, trải dài từ khoảng -20 đến 20, nhưng phần lớn các sai số nằm trong khoảng từ -10 đến 10. Điều này cho thấy mặc dù có một số dự đoán sai lệch lớn, nhưng phần lớn các sai số đều nằm trong khoảng chấp nhận được.

Đường cong mật độ KDE cho thấy phân phối gần như đối xứng với đỉnh cao nhất tại khoảng 0. Đường cong mật độ mịn, giúp xác định rõ ràng sự biến thiên của sai số dự đoán tương đối mượt mà và có hình dạng đối xứng, tương ứng với phân phối chuẩn. Điều này củng cố thêm nhận định rằng phần dư có xu hướng tuân theo phân phối chuẩn, một trong những giả định quan trọng của hồi quy tuyến tính.

Có thể kết luận rằng mô hình hồi quy đã thực hiện tốt trong việc dự đoán giá trị mục tiêu với sai số tương đối nhỏ và không có xu hướng hệ thống rõ ràng.

#### Phân Tích Phần Dư So Với Giá Trị Thực Tế



Hình 3.9 Biểu đồ phân tích phần dư so với giá trị thực tế

Đồ thị phân tán giữa giá trị thực tế và phần dư là một công cụ hữu ích để kiểm tra tính chính xác của mô hình hồi quy.

***Trục X (Giá trị thực tế):*** Biểu diễn các giá trị thực tế của biến mục tiêu trong tập dữ liệu kiểm tra (test set).

***Trục Y (Phần dư):*** Biểu diễn các phần dư (residuals) của mô hình, được tính bằng cách lấy giá trị thực tế trừ đi giá trị dự đoán (y\_test - y\_pred).

Đường ngang màu đỏ (y=0): Là đường thể hiện mức độ mà mô hình dự đoán hoàn hảo, tức là không có phần dư. Mô hình lý tưởng sẽ có các phần dư dao động quanh đường này.

Các điểm dữ liệu phân tán đều quanh đường y = 0 (đường đỏ nét đứt), cho thấy rằng sai số dự đoán không phụ thuộc vào giá trị thực tế. Đây là một dấu hiệu tốt cho thấy mô hình không bị thiên lệch (bias) theo giá trị thực tế. Không có mô hình hay xu hướng rõ ràng trong phân tán của phần dư, điều này cho thấy rằng mô hình đã bắt được hầu hết các đặc điểm của dữ liệu mà không bỏ sót hoặc quá khớp (overfitting).

Đường màu đỏ nằm ngang ở mức phần dư bằng 0 cho thấy đường trung bình của phần dư. Phần dư phân bố đều xung quanh đường này, có nghĩa là mô hình không có xu hướng hệ thống hóa việc dự đoán quá cao hoặc quá thấp.

Các điểm phần dư (các chấm màu xanh lá) phân bố khá ngẫu nhiên xung quanh đường trung bình. Điều này cho thấy rằng các lỗi dự đoán không bị ảnh hưởng bởi giá trị thực tế của biến dự đoán, tức là mô hình không có hiện tượng phương sai thay đổi (heteroskedasticity).

Giá trị thực tế dao động từ khoảng 10 đến 70, giá trị phần dư dao động từ khoảng -20 đến 20. Không có mô hình hay cấu trúc rõ ràng nào trong phân bố của các phần dư, điều này ngụ ý rằng các phần dư là ngẫu nhiên và mô hình dự đoán có thể có hiệu suất tốt.

Tóm lại, đồ thị phần dư này cho thấy rằng mô hình dự đoán có vẻ phù hợp và không có dấu hiệu của việc vi phạm các giả định cơ bản của hồi quy tuyến tính như phương sai thay đổi hay tự tương quan của các phần dư.

Từ các kết quả phân tích trên, phương trình hồi quy được viết như sau:

Y = -15119.7380 + (2.7370 \* X1 transaction date) + (-13.2411 \* X2 house age) + (-21.2341 \* X3 distance to the nearest MRT station) + (1.2256 \* X4 number of convenience stores) + (246.6449 \* X5 latitude) + (74.0661 \* X6 longitude)

### Phân tích kết quả mô hình hồi quy đa thức

#### Đánh Giá Tổng Quan Hiệu Suất Mô Hình

Giá trị hệ số R² (Coefficient of Determination) của mô hình là 0.771 điều này có nghĩa là mô hình hồi quy của tôi giải thích được 77.1% sự biến động của giá nhà trên đơn vị diện tích. Đây là một mức độ giải thích khá tốt, cho thấy rằng mô hình của tôi có khả năng dự đoán giá nhà dựa trên các biến đầu vào với một độ chính xác cao. Tuy nhiên, còn khoảng 22.9% sự biến động chưa được giải thích, có thể là do các yếu tố khác chưa được mô hình hóa hoặc các yếu tố ngẫu nhiên khác.

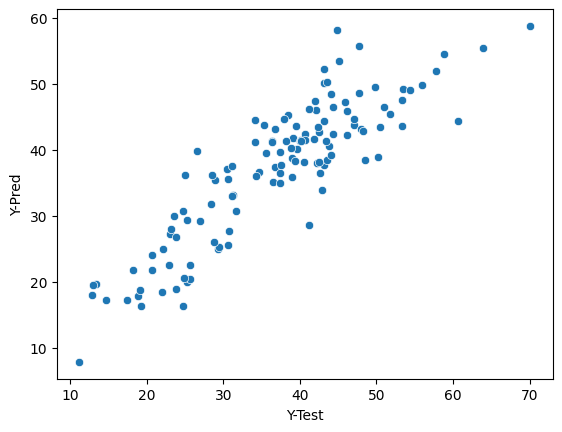
Giá trị MAE (Mean Absolute Error) là 4.478386, điều này có nghĩa là trung bình các dự đoán của mô hình sai lệch khoảng 4.478386 đơn vị so với giá trị thực tế. Giá trị MAE càng thấp, hiệu suất của mô hình càng cao. Giá trị MAE này cho thấy mức độ chính xác tương đối của mô hình.

Giá trị MSE (Root Mean Squared Error) là 30.259031, điều này có nghĩa là trung bình bình phương của sai số giữa giá trị dự đoán và giá trị thực tế là 30.259031 đơn vị. Một giá trị MSE thấp chỉ ra rằng mô hình có hiệu suất tốt. Vì sai số được bình phương, nên MSE nhạy cảm hơn với các giá trị ngoại lệ so với MAE. Giá trị MSE cao cho thấy mô hình có thể có một số sai số dự đoán lớn.

Giá trị RMSE (Root Mean Squared Error) là 5.500821 có nghĩa là trung bình sai số dự đoán của mô hình là khoảng 5.500821 đơn vị so với giá trị thực tế. RMSE nhạy cảm với các ngoại lệ và cho thấy mức độ phân tán của các sai số. Giá trị RMSE trung bình cho thấy mô hình có khả năng dự đoán khá tốt.

Các chỉ số này đều cho thấy rằng mô hình hồi quy của tôi có khả năng dự đoán tốt giá trị mục tiêu với độ chính xác ở mức trung bình. Giá trị R² cho thấy mô hình giải thích được phần lớn sự biến động của dữ liệu. Mặc dù MAE và RMSE cho thấy mức độ chính xác của mô hình hồi quy đa thức là tương đối tốt, giá trị cụ thể của các metrics này chỉ ra rằng mô hình này có thể dự đoán giá bất động sản với một sai số trung bình khoảng 4.48 đến 5.5 đơn vị. Điều này là khả thi cho các ứng dụng thực tế trong việc dự đoán giá trị bất động sản, nhưng cần cải thiện thêm để tăng độ chính xác nếu có thể

#### Phân Tích Biểu Đồ Phân Tán Giá Trị Thực Tế và Giá Trị Dự Đoán



Hình 3.10 Biểu Đồ Phân Tán Giữa Giá Trị Thực Tế và Giá Trị Dự Đoán

Biểu đồ cho thấy một xu hướng rõ ràng dọc theo đường chéo, nơi mà giá trị thực tế và giá trị dự đoán bằng nhau. Điều này cho thấy mô hình hồi quy đa thức có khả năng dự đoán tương đối tốt, vì các điểm dữ liệu nằm gần với đường chéo y=x.

***Trục X (Y-Test):*** Biểu thị giá trị thực tế của biến mục tiêu trong tập dữ liệu kiểm tra.

***Trục Y (Y-Pred):*** Biểu thị giá trị dự đoán của biến mục tiêu bởi mô hình hồi quy

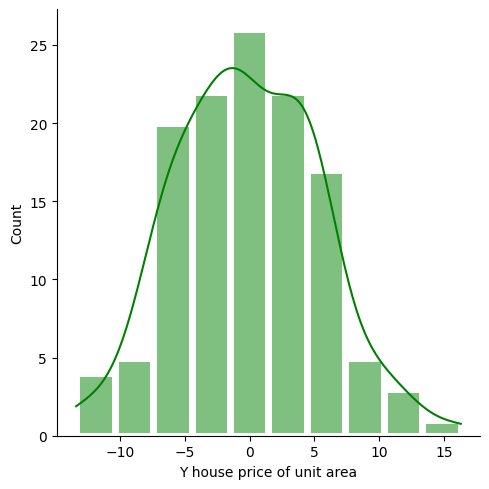
Các điểm dữ liệu khá tập trung xung quanh đường chéo, đặc biệt là ở khoảng giá trị trung bình (khoảng 20-50). Điều này chỉ ra rằng mô hình dự đoán khá chính xác cho các giá trị nằm trong khoảng này.

Tuy nhiên, một số điểm dữ liệu nằm xa khỏi đường chéo, biểu thị những dự đoán sai lệch đáng kể cho một số trường hợp cụ thể. Có thể do các yếu tố phi tuyến hoặc các biến không được bao gồm trong mô hình ảnh hưởng đến giá trị thực tế.

Ở các giá trị cao hơn của Y-Test (khoảng trên 50), có một số điểm dữ liệu cho thấy độ lệch lớn hơn so với đường chéo. Điều này có thể chỉ ra rằng mô hình không hoàn toàn chính xác khi dự đoán các giá trị cao, và có thể cần điều chỉnh thêm hoặc bổ sung thêm biến độc lập để cải thiện.

Phân bố các điểm dữ liệu không có một mẫu rõ ràng của sự phân tán không đều, mô hình không có hiện tượng overfitting nghiêm trọng đối với các giá trị trung bình, nhưng vẫn cần xem xét thêm đối với các giá trị ngoại lệ.

#### Phân Tích Phân Phối Các Phần Dư



Hình 3.11 Biểu đồ phân phối phần dư

Biểu đồ này hiển thị số lượng phần dư nằm trong mỗi khoảng giá trị và đường cong KDE mô tả phân phối xác suất của các phần dư.

***Trục X (Y house price of unit area):*** Biểu thị giá trị của phần dư.

***Trục Y (Count):*** Biểu thị số lượng các phần dư trong mỗi khoảng giá trị cụ thể.

Biểu đồ cho thấy phần dư phân bố đối xứng quanh giá trị 0 và số lượng phần dư giảm dần khi chúng xa dần khỏi trung tâm. Cho thấy rằng sai số của mô hình không bị chi phối bởi các giá trị cực đoan hoặc ngoại lệ, và phân phối của các phần dư là cân đối. Hình dạng của biểu đồ histogram gần giống với hình chuông, nghĩa là phần dư có xu hướng phân phối theo phân phối chuẩn (normal distribution). Đây là một dấu hiệu tích cực, các giả định về phân phối phần dư của mô hình hồi quy tuyến tính có thể được đáp ứng.

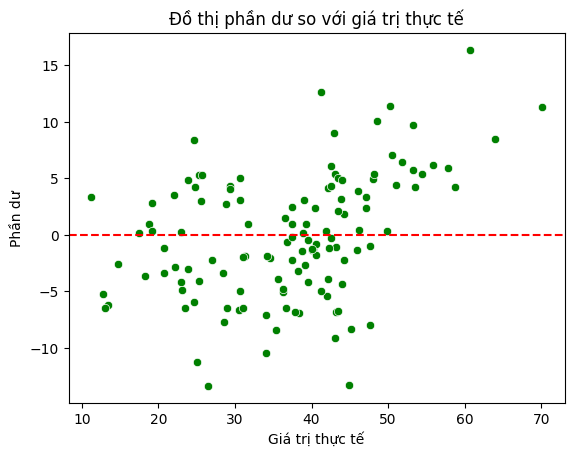
Đỉnh cao nhất của biểu đồ nằm gần giá trị 0, cho rằng hầu hết các phần dư là nhỏ và các giá trị dự đoán khá gần với giá trị thực tế. Mô hình hồi quy có khả năng dự đoán tốt với sai số nhỏ.

Đường cong mật độ KDE tương đối mượt mà và có hình dạng đối xứng, tương ứng với phân phối chuẩn. Điều này củng cố thêm nhận định rằng phần dư có xu hướng tuân theo phân phối chuẩn, một trong những giả định quan trọng của hồi quy tuyến tính.

Ngoài ra còn một số điểm dữ liệu phần dư nằm xa khỏi trung tâm (ngoại lệ) có thể được quan sát ở hai đầu của biểu đồ. Tuy nhiên, sự hiện diện của một số giá trị ngoại lệ không làm ảnh hưởng đến mô hình.

Có thể kết luận rằng mô hình hồi quy đã thực hiện tốt trong việc dự đoán giá trị mục tiêu với sai số tương đối nhỏ và không có xu hướng hệ thống rõ ràng.

#### Phân Tích Phần Dư So Với Giá Trị Thực Tế



Hình 3.12 Đồ thị phần dư so với giá trị thực tế

Đồ thị phân tán giữa giá trị thực tế và phần dư là một công cụ hữu ích để kiểm tra tính chính xác của mô hình hồi quy.

***Trục X (Giá trị thực tế):*** Biểu diễn các giá trị thực tế của biến mục tiêu trong tập dữ liệu kiểm tra (test set).

***Trục Y (Phần dư):*** Biểu diễn các phần dư (residuals) của mô hình, được tính bằng cách lấy giá trị thực tế trừ đi giá trị dự đoán (y\_test - y\_pred).

Đường ngang màu đỏ (y=0): Là đường thể hiện mức độ mà mô hình dự đoán hoàn hảo, tức là không có phần dư. Mô hình lý tưởng sẽ có các phần dư dao động quanh đường này.

Các điểm dữ liệu phần dư phân tán ngẫu nhiên quanh trục y=0. Điều này cho thấy mô hình không có xu hướng hệ thống (systematic bias) rõ ràng. Các sai số dường như phân bố đồng đều xung quanh đường 0, chỉ ra rằng mô hình không thiên lệch theo bất kỳ hướng nào cụ thể và phần dư là ngẫu nhiên.

Phần dư dương: Các điểm nằm trên đường y=0 đại diện cho các dự đoán mà mô hình đã đánh giá thấp so với giá trị thực tế. Phần dư âm: Các điểm nằm dưới đường y=0 đại diện cho các dự đoán mà mô hình đã đánh giá cao so với giá trị thực tế. Việc có cả phần dư dương và âm chỉ ra rằng mô hình đã có sự đánh giá sai lệch theo cả hai chiều, nhưng không có xu hướng quá rõ ràng trong việc chỉ đánh giá cao hoặc thấp.

Độ lan truyền tăng dần theo giá trị thực tế, một số điểm dữ liệu ở xa trục y=0, đặc biệt là khi giá trị thực tế tăng. Mô hình có thể gặp khó khăn trong việc dự đoán chính xác giá trị của các nhà có giá trị cao hoặc thấp, dẫn đến phần dư lớn hơn ở những vùng này.

Mô hình dự đoán không hoàn hảo nhưng đã có sự phân phối phần dư tương đối tốt quanh giá trị 0. Có thể cho thấy rằng mô hình có khả năng dự đoán tương đối tốt, nhưng vẫn cần cải thiện để xử lý các giá trị ngoại lệ và giảm sai số dự đoán.

Từ các kết quả phân tích trên, phương trình hồi quy được viết như sau:

Y = -147010534.4883 + (-6557.8492 \* X1 transaction date) + (8383.9599 \* X2 house age) + (118684.7824 \* X3 distance to the nearest MRT station) + (5172.8743 \* X4 number of convenience stores) + (298382.8540 \* X5 latitude) + (2357433.7148 \* X6 longitude)

### So sánh kết quả hai mô hình

Tôi tiến hành so sánh các chỉ số đánh giá hiệu suất là , MAE, MSE và RMSE để lựa chọn mô hình phù hợp nhất.

Bảng 3.5. Bảng so sánh kết quả hai mô hình

|  |  |  |
| --- | --- | --- |
| **Chỉ số** | **Hồi Quy Tuyến Tính** | **Hồi Quy Đa Thức** |
|  | 0.649 | 0.771 |
| MAE | 5.334487 | 4.478386 |
| MSE | 46.460281 | 30.259031 |
| RMSE | 6.816178 | 5.500821 |

Với R² là 0.649, mô hình hồi quy tuyến tính giải thích được 64.9% biến động của giá bất động sản, trong khi đó mô hình hồi quy đa thức với R² là 0.771 giải thích được 77.1% biến động, cho thấy mô hình này có khả năng dự báo tốt hơn.

Mô hình hồi quy đa thức có MAE thấp hơn (4.478386) so với hồi quy tuyến tính (5.334487), cho thấy rằng mô hình hồi quy đa thức có độ chính xác cao hơn trong việc dự đoán giá trị bất động sản.

MSE của mô hình hồi quy đa thức thấp hơn (30.259031) so với mô hình hồi quy tuyến tính (46.460281) cho thấy rằng mô hình hồi quy đa thức dự báo giá trị gần với thực tế hơn.

Với RMSE là 5.500821, mô hình hồi quy đa thức thể hiện khả năng dự đoán giá trị bất động sản tốt hơn so với mô hình hồi quy tuyến tính với RMSE là 6.816178.

Từ các chỉ số hiệu suất, có thể thấy rõ ràng rằng mô hình hồi quy đa thức có hiệu suất dự báo tốt hơn so với mô hình hồi quy tuyến tính trong việc dự đoán giá bất động sản tại Quận Sindian nhưng không chênh lệch nhiều.

Ngoài việc so sách các chỉ số đánh giá hiệu suất tôi cần tiến hành phân tích các hệ số hồi quy của mô hình.

***\* Mô hình hồi quy tuyến tính:***

Y = -15119.7380 + (2.7370 \* X1 transaction date) + (-13.2411 \* X2 house age) + (-21.2341 \* X3 distance to the nearest MRT station) + (1.2256 \* X4 number of convenience stores) + (246.6449 \* X5 latitude) + (74.0661 \* X6 longitude)

***X1 (transaction date):*** Hệ số dương (2.7370) gợi ý rằng các giao dịch gần đây hơn (năm lớn hơn) có giá nhà cao hơn. Điều này hợp lý vì giá bất động sản thường có xu hướng tăng lên theo thời gian do sự phát triển kinh tế và lạm phát.

***X2 (house age):*** Hệ số âm (-13.2411) cho thấy nhà càng cũ thì giá càng thấp, điều này phù hợp với thực tế vì nhà cũ thường có giá trị thấp hơn so với nhà mới.

***X3 (distance to the nearest MRT station):*** Hệ số âm (-21.2341) gợi ý rằng các căn nhà gần ga tàu điện ngầm hơn thì giá trị cao hơn, điều này hợp lý vì gần các phương tiện giao thông công cộng thường làm tăng giá trị bất động sản.

***X4 (number of convenience stores):*** Hệ số dương (1.2256) cho thấy số lượng cửa hàng tiện lợi càng nhiều, giá nhà càng cao, điều này hợp lý vì các tiện ích gần nhà thường làm tăng giá trị bất động sản.

***X5 (latitude) và X6 (longitude):*** Các hệ số này (246.6449 và 74.0661) chỉ ra rằng vị trí địa lý (vĩ độ và kinh độ) có tác động đến giá nhà. Sự khác biệt về vĩ độ và kinh độ có thể đại diện cho sự khác biệt về vị trí địa lý và mức độ phát triển của các khu vực, làm ảnh hưởng đến giá bất động sản.

***\* Hồi quy đa thức:***

Phương trình hồi quy đa thức là:

Y = -55.3433 + (1.4919 \* X1 transaction date) + (-37.0560 \* X2 house age) + (89.5931 \* X3 distance to the nearest MRT station) + (74.8746 \* X4 number of convenience stores) + (101.8644 \* X5 latitude) + (141.3042 \* X6 longitude) + (4.5632 \* X1 transaction date^2) + (-0.3529 \* X1 transaction date X2 house age) + (-0.7836 \* X1 transaction date X3 distance to the nearest MRT station) + (-1.9644 \* X1 transaction date X4 number of convenience stores) + (-16.6411 \* X1 transaction date X5 latitude) + (8.8463 \* X1 transaction date X6 longitude) + (41.6364 \* X2 house age^2) + (6.9902 \* X2 house age X3 distance to the nearest MRT station) + (-3.2417 \* X2 house age X4 number of convenience stores) + (-27.8541 \* X2 house age X5 latitude) + (-0.0041 \* X2 house age X6 longitude) + (-8.1235 \* X3 distance to the nearest MRT station^2) + (-78.4613 \* X3 distance to the nearest MRT station X4 number of convenience stores) + (-125.9188 \* X3 distance to the nearest MRT station X5 latitude) + (-61.4973 \* X3 distance to the nearest MRT station X6 longitude) + (-8.0979 \* X4 number of convenience stores^2) + (-76.6776 \* X4 number of convenience stores X5 latitude) + (-21.7486 \* X4 number of convenience stores X6 longitude) + (9.1669 \* X5 latitude^2) + (-22.9495 \* X5 latitude X6 longitude) + (-80.7561 \* X6 longitude^2)

**Giải Thích Các Hệ Số**

**Hệ số độc lập (Intercept):**

***−55.3433***

Đây là giá trị của Y khi tất cả các biến đầu vào đều bằng 0. Mặc dù có thể không có ý nghĩa trực tiếp trong thực tế vì các biến đầu vào không bao giờ bằng 0, nhưng nó giúp thiết lập điểm bắt đầu cho phương trình.

**Hệ số tuyến tính (Linear Coefficients):**

***1.4919⋅X1***

Mỗi đơn vị tăng lên trong ngày giao dịch (X1), giá trị Y sẽ tăng thêm 1.4919 đơn vị, giả sử các biến khác không đổi**.**

***−37.0560⋅X2***

Mỗi đơn vị tăng lên trong tuổi nhà (X2), giá trị Y sẽ giảm 37.0560 đơn vị, giả sử các biến khác không đổi***.***

***89.5931⋅X3***

Mỗi đơn vị tăng lên trong khoảng cách đến ga MRT (X3), giá trị Y sẽ tăng thêm 89.5931 đơn vị, giả sử các biến khác không đổi.

***74.8746⋅X4***

Mỗi đơn vị tăng lên trong số cửa hàng tiện lợi (X4), giá trị Y sẽ tăng thêm 74.8746 đơn vị, giả sử các biến khác không đổi.

***101.8644⋅X5***

Mỗi đơn vị tăng lên trong vĩ độ (X5), giá trị Y sẽ tăng thêm 101.8644 đơn vị, giả sử các biến khác không đổi.

***141.3042⋅X6***

Mỗi đơn vị tăng lên trong kinh độ (X6), giá trị Y sẽ tăng thêm 141.3042 đơn vị, giả sử các biến khác không đổi.

**Hệ số bậc hai (Quadratic Coefficients):**

***4.5632⋅X1^2***

Mỗi đơn vị tăng lên trong bình phương của ngày giao dịch (X1^2), giá trị Y sẽ tăng thêm 4.5632 đơn vị.

***41.6364⋅X2^2***

Mỗi đơn vị tăng lên trong bình phương của tuổi nhà (X2^2), giá trị Y sẽ tăng thêm 41.6364 đơn vị.

***−8.1235⋅X3^2***

Mỗi đơn vị tăng lên trong bình phương của khoảng cách đến ga MRT (X3^2), giá trị Y sẽ giảm 8.1235 đơn vị.

***−8.0979⋅X4^2***

Mỗi đơn vị tăng lên trong bình phương của số cửa hàng tiện lợi (X4^2), giá trị Y sẽ giảm 8.0979 đơn vị.

***9.1669⋅X5^2***

Mỗi đơn vị tăng lên trong bình phương của vĩ độ (X5^2), giá trị Y sẽ tăng thêm 9.1669 đơn vị.

***−80.7561⋅X6^2***

Mỗi đơn vị tăng lên trong bình phương của kinh độ (X6^2), giá trị Y sẽ giảm 80.7561 đơn vị.

**Hệ số tương tác (Interaction Coefficients):**

***−0.3529⋅X1⋅X2***

Mỗi đơn vị tăng lên trong sản phẩm của ngày giao dịch và tuổi nhà, giá trị Y sẽ giảm 0.3529 đơn vị***.***

***−0.7836⋅X1⋅X3***

Mỗi đơn vị tăng lên trong sản phẩm của ngày giao dịch và khoảng cách đến ga MRT, giá trị Y sẽ giảm 0.7836 đơn vị.

***−1.9644⋅X1⋅X4***

Mỗi đơn vị tăng lên trong sản phẩm của ngày giao dịch và số cửa hàng tiện lợi, giá trị Y sẽ giảm 1.9644 đơn vị.

***−16.6411⋅X1⋅X5***

Mỗi đơn vị tăng lên trong sản phẩm của ngày giao dịch và vĩ độ, giá trị Y sẽ giảm 16.6411 đơn vị.

***8.8463⋅X1⋅X6***

Mỗi đơn vị tăng lên trong sản phẩm của ngày giao dịch và kinh độ, giá trị Y sẽ tăng 8.8463 đơn vị.

***6.9902⋅X2⋅X3***

Mỗi đơn vị tăng lên trong sản phẩm của tuổi nhà và khoảng cách đến ga MRT, giá trị Y sẽ tăng 6.9902 đơn vị.

***−3.2417⋅X2⋅X4***

Mỗi đơn vị tăng lên trong sản phẩm của tuổi nhà và số cửa hàng tiện lợi, giá trị Y sẽ giảm 3.2417 đơn vị.

***−27.8541⋅X2⋅X5***

Mỗi đơn vị tăng lên trong sản phẩm của tuổi nhà và vĩ độ, giá trị Y sẽ giảm 27.8541 đơn vị.

***−0.0041⋅X2⋅X6***

Mỗi đơn vị tăng lên trong sản phẩm của tuổi nhà và kinh độ, giá trị Y sẽ giảm 0.0041 đơn vị.

***−78.4613⋅X3⋅X4***

Mỗi đơn vị tăng lên trong sản phẩm của khoảng cách đến ga MRT và số cửa hàng tiện lợi, giá trị Y sẽ giảm 78.4613 đơn vị.

***−125.9188⋅X3⋅X5***

Mỗi đơn vị tăng lên trong sản phẩm của khoảng cách đến ga MRT và vĩ độ, giá trị Y sẽ giảm 125.9188 đơn vị.

***−61.4973⋅X3⋅X6***

Mỗi đơn vị tăng lên trong sản phẩm của khoảng cách đến ga MRT và kinh độ, giá trị Y sẽ giảm 61.4973 đơn vị.

***−76.6776⋅X4⋅X5***

Mỗi đơn vị tăng lên trong sản phẩm của số cửa hàng tiện lợi và vĩ độ, giá trị Y sẽ giảm 76.6776 đơn vị.

***−21.7486⋅X4⋅X6***

Mỗi đơn vị tăng lên trong sản phẩm của số cửa hàng tiện lợi và kinh độ, giá trị Y sẽ giảm 21.7486 đơn vị.

***−22.9495⋅X5⋅X6***

Mỗi đơn vị tăng lên trong sản phẩm của vĩ độ và kinh độ, giá trị Y sẽ giảm 22.9495 đơn vị.

Dựa trên các chỉ số đánh giá và hệ số hồi quy của hai mô hình. Mặc dù các chỉ số đánh giá của mô hình hồi quy đa thức cho thấy nó có hiệu suất dự báo tốt hơn mô hình hồi quy tuyến tính nhưng sự chênh lệch là không lớn. Trong khi đó, các hệ số hồi quy của mô hình hồi quy tuyến tính phù hợp với các lý thuyết kinh tế hơn mô hình hồi quy đa thức. Vì thế, tôi kết luận rằng, mô hình hồi quy tuyến tính là lựa chọn tốt hơn để dự báo giá bất động sản trong nghiên cứu này.

## Ứng dụng kết quả mô hình hồi quy tuyến tính

Tôi sử dụng mô hình hồi quy tuyến tính đã huấn luyện để tiến hành dự đoán giá bất động sản cho các căn nhà tại Quận Sindian. Một ví dụ cụ thể là dự báo giá cho căn nhà với các đặc điểm sau:

Thời gian giao dịch: 2023

Tuổi nhà: 10 năm

Khoảng cách đến ga tàu điện ngầm gần nhất: 250 mét

Số lượng cửa hàng tiện lợi: 5

Vĩ độ: 24.972

Kinh độ: 121.541

Kết quả dự báo cho thấy giá trị bất động sản của căn nhà này là khoảng 143.74 trên mỗi đơn vị diện tích. Tôi đã thực hiện một số kịch bản giả định để phân tích ảnh hưởng của các yếu tố độc lập lên giá bất động sản. Dưới đây là một số ví dụ:

***\* Thay Đổi Tuổi Nhà:***

Giả sử tuổi nhà tăng từ 10 lên 20 năm, giữ nguyên các yếu tố khác. Kết quả cho thấy giá nhà có xu hướng giảm mạnh (giảm 132.41 đơn vị), phản ánh rằng nhà cũ hơn thường có giá trị thấp hơn do chi phí bảo trì và cải tạo cao hơn.

***\* Thay Đổi Khoảng Cách Đến Ga Tàu Điện Ngầm*:**

Khi khoảng cách từ nhà đến ga tàu điện ngầm tăng từ 250 mét lên 255 mét, giá nhà giảm đáng kể (giảm 106.17 đơn vị). Điều này cho thấy việc gần gũi với phương tiện giao thông công cộng có tác động tích cực đến giá trị bất động sản.

***\* Tăng Số Lượng Cửa Hàng Tiện Lợi Gần Kề:***

Nếu số lượng cửa hàng tiện lợi gần căn nhà tăng từ 5 lên 10, giá trị nhà có xu hướng tăng lên (tăng 6.12). Điều này minh chứng rằng các tiện ích và dịch vụ xung quanh đóng vai trò quan trọng trong việc tăng giá trị của bất động sản.

Kết quả từ các phân tích tình huống giả định này cung cấp cái nhìn sâu sắc về cách mà các yếu tố độc lập có thể ảnh hưởng đến giá bất động sản. Điều này không chỉ giúp hiểu rõ hơn về thị trường bất động sản tại Quận Sindian mà còn cung cấp thông tin hữu ích cho các nhà đầu tư và người mua nhà trong việc ra quyết định.

# KẾT LUẬN VÀ KIẾN NGHỊ

Dựa trên kết quả của mô hình hồi quy tuyến tính, tôi nhận thấy rằng mô hình này đã đạt được một mức độ dự đoán hợp lý về giá bất động sản tại Quận Sindian, thành phố Tân Bắc, Đài Loan, thể hiện qua hệ số R² là 0.649, cùng với các chỉ số lỗi dự đoán như MAE, MSE và RMSE ở mức chấp nhận được. Mặc dù mô hình hồi quy tuyến tính có hiệu suất kém hơn so với mô hình hồi quy đa thức, nó vẫn cung cấp một cái nhìn tổng quát về các yếu tố ảnh hưởng đến giá bất động sản, cho thấy sự đơn giản và tính dễ hiểu của mô hình. Các kết quả dự đoán cho thấy rằng mô hình này có thể nắm bắt được phần lớn các biến động trong dữ liệu bất động sản, mặc dù còn tồn tại một số sai số dự đoán cần được cải thiện.

Để nâng cao hiệu suất của mô hình, tôi khuyến nghị cần phải bổ sung thêm dữ liệu và xem xét các yếu tố khác có thể ảnh hưởng đến giá bất động sản như các yếu tố kinh tế vĩ mô và sự phát triển hạ tầng. Ngoài ra, việc thường xuyên cập nhật và hiệu chỉnh mô hình sẽ giúp nâng cao độ chính xác và tính phản ánh thực tế của mô hình trước những biến động của thị trường. Các công ty bất động sản và nhà đầu tư có thể sử dụng mô hình này như một công cụ tham khảo để đưa ra các quyết định kinh doanh phù hợp, đồng thời tiếp tục nghiên cứu và phát triển các mô hình phân tích nâng cao nhằm tối ưu hóa khả năng dự báo và giảm thiểu rủi ro trong các giao dịch bất động sản.

# TÀI LIỆU THAM KHẢO

Barr, D. (2018, June 28). *Real estate valuation using regression analysis – A tutorial*. Toptal Finance Blog; Toptal.

Constantinescu, M. (2019, March 14). *Machine-Learning real estate valuation: Not only a data affair*. Towards Data Science.

*Real estate valuation Using linear regression*. (2021, September 12). Kaggle.com; Kaggle.

Ozgur, C. (n.d.). *Multiple linear regression applications in real estate pricing*.

(N.d.). Amazon.com

*Xử lý các giá trị ngoại lệ — Machine Learning cho dữ liệu dạng bảng*. (n.d.). Machinelearningcoban.com.

*Bài mẫu tiểu luận về kinh tế lượng, HAY, 9 ĐIỂM*. (n.d.)

*Xây dựng mô hình định giá bất động sản tự động hỗ trợ quá trình thẩm định giá trị tài sản bảo đảm*. (n.d.). Gov.vn.

Phạm Văn, T. (2016, November 24). *Mô hình hồi quy ứng dụng trong bài toán dự đoán giá bất động sản - Machine Learning (phần 2)*. Viblo.

Tung N. D. (2017, December 27). *Machine Learning thật thú vị (1): Dự đoán giá nhà đất*. Spiderum.

Phạm Văn, T. (2016b, November 28). *Mô hình hồi quy ứng dụng trong bài toán dự đoán giá bất động sản - Machine Learning (phần 3)*. Viblo.