ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC SÀI GÒN

**KHOA CÔNG NGHỆ THÔNG TIN**





# **Khảo sát bài báo cho dữ liệu Ames Housing**

**Giảng viên hướng dẫn:** Đỗ Như Tài

**Sinh viên thực hiện:** Đặng Huy Hoàng – 3122560019

Đặng Huy Hoàng – 3122560020

**TP. HỒ CHÍ MINH, tháng 3 năm 2025**

[**Khảo sát bài báo cho dữ liệu Ames Housing 1**](#_78nb8wh3ovp5)

[**1. Giới thiệu 3**](#_cxgjoqzcms3m)

[1.1. Bối cảnh và tầm quan trọng: 3](#_acumqqets26r)

[1.3. Phương pháp tiếp cận: 4](#_3yd9xnyd1yf)

[1.4. Mục tiêu và đóng góp: 4](#_to6w9caty7j3)

[**2. Nguồn gốc và thông tin chung 4**](#_x5fyxbcipish)

[2.1. Nguồn gốc dữ liệu và bối cảnh: 4](#_wxvtmwc2ozb7)

[2.3. Đặc điểm của dữ liệu: 5](#_xr3jedmouyug)

[**3. Từ điễn dữ liệu 6**](#_j8yre0h6qs0z)

[3.1. Phân loại biến số: 6](#_tcbzw7y37tcx)

[3.2. Các đặc điểm quan trọng và ý nghĩa: 6](#_vgeyojk2bij7)

[3.3. Phân nhóm đặc điểm: 7](#_3tsu6ywomqsw)

[**4. Các Bài Toán Liên Quan 7**](#_rydbul20yg8t)

[4.1. Hai bài toán trọng tâm: 7](#_y642x35mpxf2)

[4.2. Phương pháp tiếp cận cho bài toán 1: Xác định các yếu tố quan trọng 8](#_71wpxwlxhzxc)

[4.3. Phương pháp tiếp cận cho bài toán 2: Dự đoán giá nhà 9](#_41vu9hm0e04j)

[4.4. Đánh giá hiệu suất mô hình: 9](#_i3czaz6ddg4q)

[**5. Kết quả đạt được và độ đo 9**](#_5ylwvk66pxza)

[5.1. Lựa chọn mô hình tối ưu: 9](#_j29bp62wkw1y)

[5.2. Kiểm định chéo (Cross-Validation): 11](#_9nt5f0r3qyvw)

[**6. Tính ứng dụng thực tế 12**](#_57vn1wp3lfle)

[6.1. Hỗ trợ người mua nhà đưa ra quyết định sáng suốt: 12](#_ylnr2lkxnewj)

[6.2. Cung cấp công cụ hỗ trợ ra quyết định cho nhà đầu tư bất động sản: 12](#_40usrfnhgph)

[6.3. Hỗ trợ hoạch định chính sách và quản lý thị trường: 12](#_9by8y0cugjrj)

[**7. Các bài khảo sát liên quan 13**](#_t492ude44lmp)

[7.1. Rahidi et al. (2015): Phân Loại Các Yếu Tố Ảnh Hưởng 13](#_zdn9qpn8pdfg)

[7.2. Zhou (2021): Kết Hợp Hồi Quy Đa Thức và Tối Ưu Hóa 13](#_cewq6t5yunlc)

[7.3. Alfiyatin et al. (2017): Một Nghiên Cứu Tương Tự 14](#_62iq7biwuvwi)

[**8. Kết Luận 14**](#_fgtwubo8j3f)

[8.1. Thành công và đóng góp: 15](#_64s5ukvf8o8a)

[8.2. Hạn chế và thách thức: 15](#_odi8up9kvhf)

[8.3. Hướng phát triển tương lai: 16](#_qep4ncrrdka3)

[**9. Tài Liệu Tham Khảo 16**](#_vmj6b6wzq6w)

[[1] D. De Cock, "Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project," Journal of Statistics Education, vol. 19, no. 3, 2011. 17](#_nw7l6hcjdse2)

[[2] R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartoto, and I. B. Syamwil, "Factors influencing the price of housing in Indonesia," International Journal of Housing Market Analysis, 2015. 17](#_9v2mq7ove4m6)

[[3] C. Zhou, "House price prediction using polynomial regression with particle swarm optimization," Journal of Physics: Conference Series, 2021. 17](#_2gxkyhde81k)

[[4] A. N. Alfiyatin, H. Taufiq, R. E. Febrita, and W. F. Mahmudy, "Modeling house price prediction using regression analysis and particle swarm optimization," International Journal of Advanced Computer Science and Applications, 2017. 17](#_jt6kt0wdn5ax)

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

# **Giới thiệu**

## **Bối cảnh và tầm quan trọng:**

* Trong bối cảnh đô thị hóa nhanh chóng và sự gia tăng dân số tại Hoa Kỳ, giá nhà đất liên tục biến động và trở thành một trong những thách thức lớn. Việc dự đoán giá nhà không chỉ là mối quan tâm của người mua và nhà đầu tư, mà còn ảnh hưởng trực tiếp đến sự ổn định của thị trường bất động sản và nền kinh tế nói chung. Sự biến động giá nhà có thể tác động đến niềm tin của người tiêu dùng, quyết định chi tiêu, và thậm chí có thể dẫn đến suy thoái kinh tế. Do đó, khả năng dự báo chính xác giá nhà trở thành một yếu tố then chốt, giúp các bên liên quan đưa ra quyết định sáng suốt và giảm thiểu rủi ro.
  1. **Vấn đề nghiên cứu:**
* Bài báo "Price Prediction of Ames Housing Through Advanced Regression Techniques" của Yueting Han giải quyết trực tiếp vấn đề này bằng cách tập trung vào thị trường nhà đất cụ thể tại Ames, Iowa. Bài báo không chỉ dừng lại ở việc dự đoán giá nhà, mà còn đi sâu vào việc xác định các yếu tố quan trọng ảnh hưởng đến giá trị bất động sản. Điều này đặc biệt quan trọng vì không phải ai cũng là chuyên gia trong lĩnh vực bất động sản, và việc hiểu rõ các yếu tố này giúp người mua và nhà đầu tư đưa ra quyết định thông minh hơn.

## **Phương pháp tiếp cận:**

* Để đạt được mục tiêu này, tác giả Yueting Han sử dụng một loạt các kỹ thuật hồi quy tiên tiến, kết hợp với các phương pháp học máy (machine learning). Các kỹ thuật này bao gồm phân tích tương quan (correlation), kỹ thuật đặc trưng (feature engineering), và xây dựng mô hình (model building). Dữ liệu được sử dụng là một bộ dữ liệu toàn diện, bao gồm hầu hết các đặc điểm của nhà ở tại Ames, Iowa. Các mô hình được lựa chọn để đánh giá bao gồm LASSO Regression, Elastic Net Regression, Gradient Boosting Regression, XGBoost, LightGBM, và Stacked Model. Hiệu suất của các mô hình được đánh giá thông qua phương pháp kiểm chứng chéo (cross-validation) và chỉ số sai số trung bình bình phương gốc (root-mean-square logarithmic error).

## **Mục tiêu và đóng góp:**

* Bài báo này có hai mục tiêu chính. Thứ nhất, xác định các đặc trưng (features) thực sự liên quan đến sự biến động của giá nhà. Thứ hai, xây dựng các mô hình dự đoán giá trị của từng ngôi nhà trong tập dữ liệu kiểm tra (test set). Bằng cách kết hợp các kỹ thuật thống kê và học máy, bài báo không chỉ cung cấp một phương pháp dự đoán giá nhà hiệu quả, mà còn làm sáng tỏ các yếu tố quan trọng ảnh hưởng đến giá trị bất động sản, từ đó mang lại lợi ích cho người mua, nhà đầu tư và toàn xã hội.

# **Nguồn gốc và thông tin chung**

## **Nguồn gốc dữ liệu và bối cảnh:**

* Cơ sở dữ liệu nền tảng cho bài báo "Price Prediction of Ames Housing Through Advanced Regression Techniques" là bộ dữ liệu "Ames Housing dataset", một tập dữ liệu nổi tiếng và được sử dụng rộng rãi trong cộng đồng khoa học dữ liệu và các cuộc thi về học máy. Mặc dù bài báo của Yueting Han không trực tiếp trích dẫn nguồn gốc ban đầu, dữ liệu này thường được liên kết chặt chẽ với công trình nghiên cứu của Dean De Cock vào năm 2011. De Cock [[1](#_nw7l6hcjdse2)], khi đó là giảng viên tại Đại học Truman State, đã thu thập và chuẩn hóa dữ liệu này từ hồ sơ của Văn phòng Đánh giá Bất động sản Thành phố Ames (Ames Assessor's Office). Mục đích ban đầu của De Cock là tạo ra một bộ dữ liệu thực tế, phong phú và phức tạp hơn để thay thế cho bộ dữ liệu Boston Housing dataset vốn đã trở nên lỗi thời và đơn giản trong giảng dạy thống kê.
  1. **Thông tin chung về bộ dữ liệu**
* Bộ dữ liệu Ames Housing trở nên phổ biến thông qua cuộc thi "House Prices: Advanced Regression Techniques" trên nền tảng Kaggle, một cộng đồng trực tuyến dành cho các nhà khoa học dữ liệu và những người đam mê học máy. Cuộc thi này đã thu hút sự tham gia của hàng ngàn người, tạo ra một môi trường cạnh tranh và hợp tác để phát triển các mô hình dự đoán giá nhà tốt nhất.
* Bộ dữ liệu có thể được tải xuống trực tiếp từ trang web của cuộc thi trên Kaggle(<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques>). Để phục vụ cho mục đích nghiên cứu và phát triển mô hình, dữ liệu được chia thành hai tập riêng biệt:
* **Tập huấn luyện (train set):** Tập này bao gồm thông tin chi tiết của 1460 ngôi nhà ở Ames, Iowa. Mỗi ngôi nhà được mô tả bằng 81 đặc trưng (features), bao gồm các thông tin về kích thước, chất lượng, vị trí, tiện nghi, và các yếu tố khác có thể ảnh hưởng đến giá trị. Quan trọng nhất, tập huấn luyện chứa biến mục tiêu (target variable) là 'SalePrice' – giá bán thực tế của ngôi nhà. Thông tin này được sử dụng để "dạy" các mô hình học máy, giúp chúng tìm ra mối quan hệ giữa các đặc trưng và giá bán.
* **Tập kiểm tra (test set):** Tập này chứa thông tin của 1459 ngôi nhà khác, cũng với 80 đặc trưng tương tự như tập huấn luyện. Tuy nhiên, điểm khác biệt quan trọng là tập kiểm tra không bao gồm giá bán thực tế ('SalePrice'). Mục đích của tập này là để đánh giá hiệu suất của các mô hình đã được huấn luyện. Các mô hình sẽ dự đoán giá bán cho các ngôi nhà trong tập kiểm tra, và kết quả dự đoán sẽ được so sánh với giá bán thực tế (mà chỉ ban tổ chức cuộc thi biết) để xác định độ chính xác.

## **Đặc điểm của dữ liệu:**

* Các đặc trưng trong bộ dữ liệu Ames Housing rất đa dạng, bao gồm cả biến số (numerical) và biến phân loại (categorical). Chúng mô tả các khía cạnh khác nhau của ngôi nhà, từ các yếu tố vật lý như diện tích, số phòng ngủ, năm xây dựng, đến các yếu tố định tính như chất lượng tổng thể, loại vật liệu, khu vực lân cận, và các tiện ích khác. Sự phong phú và đa dạng của dữ liệu này tạo ra một thách thức thú vị cho việc xây dựng mô hình dự đoán, đồng thời cũng cung cấp một cơ hội để khám phá các yếu tố quan trọng ảnh hưởng đến giá nhà.

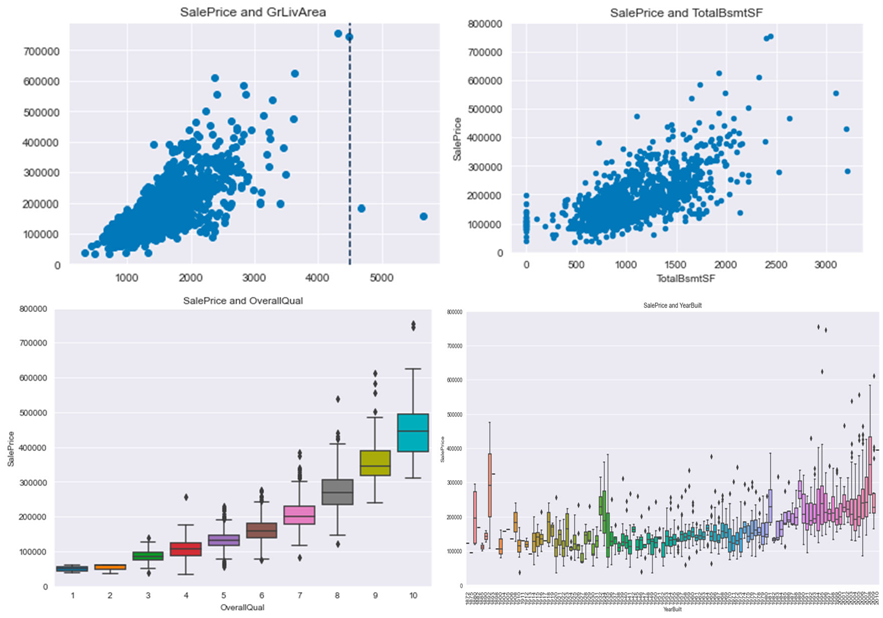
# **Từ điễn dữ liệu**

## **Phân loại biến số:**

* Để hiểu rõ hơn về các yếu tố tác động đến giá nhà trong bộ dữ liệu Ames Housing, cần xem xét kỹ lưỡng các loại biến số và ý nghĩa của chúng. Dữ liệu được cấu trúc thành hai loại biến chính:
* **Biến số (Numerical Variables):** Đây là các biến thể hiện bằng các giá trị số, có thể đo lường được. Ví dụ, diện tích, số phòng, năm xây dựng, v.v. Các biến số này cung cấp thông tin định lượng về các đặc điểm của ngôi nhà.
* **Biến phân loại (Categorical Variables):** Đây là các biến thể hiện bằng các giá trị không phải số, thường là các nhãn hoặc tên gọi, mô tả các đặc điểm định tính của ngôi nhà. Ví dụ, loại vật liệu xây dựng, khu vực lân cận, kiểu mái nhà, v.v. Các biến này cung cấp thông tin về loại, phong cách, hoặc đặc điểm chất lượng của ngôi nhà.

## **Các đặc điểm quan trọng và ý nghĩa:**

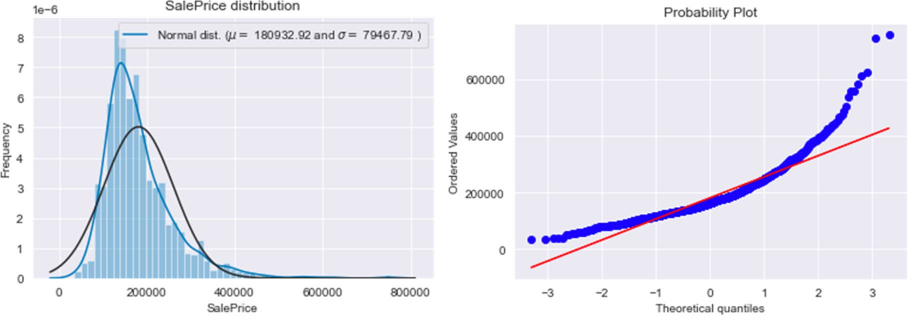
* Bài báo của Yueting Han nhấn mạnh một số đặc điểm quan trọng có ảnh hưởng đáng kể đến giá nhà. Việc hiểu rõ ý nghĩa của các đặc điểm này giúp ta hình dung rõ hơn về cách chúng tác động đến giá trị bất động sản:
* **'OverallQual' (Overall Quality):** Đây là một biến phân loại, thể hiện chất lượng tổng thể của vật liệu và hoàn thiện của ngôi nhà. Biến này được đánh giá trên thang điểm từ 1 (rất kém) đến 10 (xuất sắc). 'OverallQual' được xem là một trong những yếu tố quan trọng nhất, vì nó phản ánh trực tiếp chất lượng và sự hấp dẫn của ngôi nhà đối với người mua.
* **'YearBuilt' (Year Built):** Đây là một biến số, cho biết năm ngôi nhà được xây dựng ban đầu. Thông thường, những ngôi nhà mới hơn có xu hướng có giá cao hơn do sử dụng vật liệu xây dựng hiện đại, thiết kế hợp thời, và ít cần sửa chữa hơn.
* **'TotalBsmtSF' (Total Basement Surface Area):** Đây là một biến số, thể hiện tổng diện tích tầng hầm (tính bằng square feet). Tầng hầm rộng rãi có thể cung cấp thêm không gian sử dụng cho gia đình, như phòng giải trí, phòng tập thể dục, hoặc không gian lưu trữ, do đó làm tăng giá trị của ngôi nhà.
* **'GrLivArea' (Gross Living Area):** Đây là một biến số, chỉ diện tích sinh hoạt trên mặt đất (tính bằng square feet), không bao gồm tầng hầm. 'GrLivArea' là một yếu tố then chốt khác, vì nó phản ánh trực tiếp không gian sống mà gia đình có thể sử dụng hàng ngày.
* **'Neighborhood' (Neighborhood):** Đây là một biến phân loại, mô tả vị trí của ngôi nhà trong các khu vực lân cận khác nhau của thành phố Ames. Vị trí địa lý đóng vai trò quan trọng trong việc định giá nhà, vì nó ảnh hưởng đến nhiều yếu tố như sự an toàn, sự yên tĩnh, cảnh quan, và khả năng tiếp cận các tiện ích công cộng (trường học, bệnh viện, công viên, trung tâm mua sắm, v.v.).

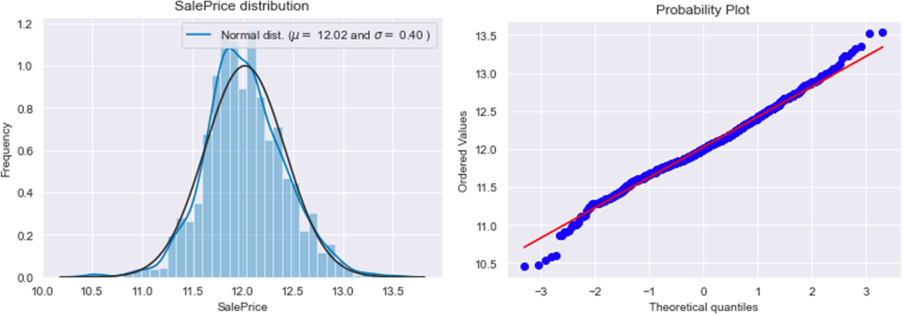


**Figure 1.** Part of visualization between variables and‘SalePrice’

## **Phân nhóm đặc điểm:**

* Để có cái nhìn tổng quan và hệ thống hơn, các đặc điểm trong bộ dữ liệu Ames Housing có thể được phân loại thành ba nhóm chính:
* **'Building' (Đặc điểm xây dựng):** Nhóm này bao gồm các đặc điểm liên quan đến cấu trúc và chất lượng vật lý của ngôi nhà. Ví dụ: 'OverallQual', 'YearBuilt', loại vật liệu xây dựng, kiểu mái nhà, v.v.
* **'Space' (Đặc điểm không gian):** Nhóm này tập trung vào các đặc điểm liên quan đến diện tích và không gian sử dụng của ngôi nhà. Ví dụ: 'TotalBsmtSF', 'GrLivArea', số phòng ngủ, số phòng tắm, v.v.
* **'Location' (Đặc điểm vị trí):** Nhóm này bao gồm các đặc điểm liên quan đến vị trí địa lý và môi trường xung quanh của ngôi nhà. Ví dụ: 'Neighborhood', khoảng cách đến các tiện ích công cộng, cảnh quan, v.v.





**Figure 2.** Plots of‘SalePrice’ before and after log transformation

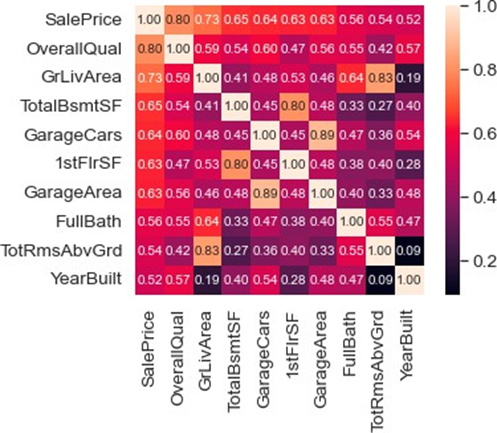
# **Các Bài Toán Liên Quan**

## **Hai bài toán trọng tâm:**

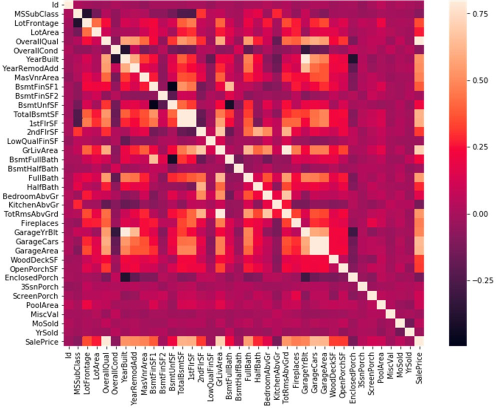
* Bài báo "Price Prediction of Ames Housing Through Advanced Regression Techniques" của Yueting Han không chỉ đơn thuần là một bài tập dự đoán giá nhà. Nó tập trung giải quyết hai bài toán cốt lõi, có mối quan hệ chặt chẽ với nhau trong lĩnh vực bất động sản:
* **Xác định các yếu tố then chốt ảnh hưởng đến giá trị nhà:** Đây là bài toán mang tính khám phá và giải thích. Nó không chỉ dừng lại ở việc dự đoán, mà còn đi sâu vào việc tìm hiểu nguyên nhân, xác định các đặc điểm (features) nào của ngôi nhà thực sự có tác động đáng kể đến giá bán. Bài toán này có ý nghĩa quan trọng đối với cả người mua và người bán, giúp họ hiểu rõ hơn về các yếu tố cần quan tâm khi đánh giá một bất động sản.
* **Dự đoán giá nhà một cách chính xác:** Đây là bài toán mang tính ứng dụng cao. Sau khi đã xác định được các yếu tố quan trọng, bài toán tiếp theo là xây dựng một mô hình có khả năng dự đoán giá nhà dựa trên các yếu tố đó. Mô hình này cần phải đạt được độ chính xác cao, để có thể trở thành một công cụ hữu ích cho các bên liên quan trong thị trường bất động sản, như người mua, người bán, nhà đầu tư, và các chuyên gia định giá.

## **Phương pháp tiếp cận cho bài toán 1: Xác định các yếu tố quan trọng**

* Để giải quyết bài toán thứ nhất, tác giả sử dụng kết hợp các phương pháp thống kê và kỹ thuật học máy:
* **Phân tích tương quan (Correlation Analysis):** Phương pháp này được sử dụng để đo lường mức độ quan hệ tuyến tính giữa các đặc điểm của ngôi nhà và giá bán. Cụ thể, bài báo sử dụng hệ số tương quan Pearson để xác định các đặc điểm có tương quan mạnh nhất với 'SalePrice'. Ví dụ, kết quả phân tích cho thấy 'OverallQual' (chất lượng tổng thể) có hệ số tương quan cao nhất (0.8), cho thấy mối quan hệ rất chặt chẽ giữa chất lượng và giá nhà. Tương tự, 'GrLivArea' (diện tích sinh hoạt) và 'TotalBsmtSF' (diện tích tầng hầm) cũng có tương quan đáng kể (lần lượt là 0.73 và 0.65).
* **Kỹ thuật chọn lọc đặc trưng (Feature Engineering):** Đây là quá trình tạo ra các đặc trưng mới, hoặc biến đổi các đặc trưng hiện có, để cải thiện hiệu suất của mô hình dự đoán. Nó cũng bao gồm việc loại bỏ các đặc trưng không cần thiết hoặc gây nhiễu. Ví dụ, bài báo đã loại bỏ các đặc trưng có nhiều giá trị thiếu (missing values), hoặc các đặc trưng có tương quan quá cao với nhau (multicollinearity) để tránh làm giảm độ chính xác của mô hình.



**Figure 3.** Heatmap of the top 10 most correlated variablesto ‘SalePrice’



**Figure 4.**.Heatmap of Correlations Between Variables

## **Phương pháp tiếp cận cho bài toán 2: Dự đoán giá nhà**

* Bài toán thứ hai được giải quyết bằng cách sử dụng một loạt các mô hình hồi quy tuyến tính và phi tuyến:
* **LASSO Regression và Elastic Net Regression:** Đây là hai mô hình hồi quy tuyến tính có thêm các thành phần điều chuẩn (regularization) để tránh hiện tượng quá khớp (overfitting) và giảm độ phức tạp của mô hình.
* **Gradient Boosting Regression, XGBoost, và LightGBM:** Đây là các mô hình dựa trên cây quyết định (decision tree) sử dụng kỹ thuật boosting, kết hợp nhiều cây quyết định yếu để tạo thành một mô hình mạnh mẽ và chính xác hơn.
* **Stacked Model:** Đây là một kỹ thuật kết hợp (ensemble technique) sử dụng nhiều mô hình khác nhau để đưa ra dự đoán cuối cùng. Ý tưởng là kết hợp điểm mạnh của các mô hình khác nhau để đạt được kết quả tốt hơn so với bất kỳ mô hình đơn lẻ nào.

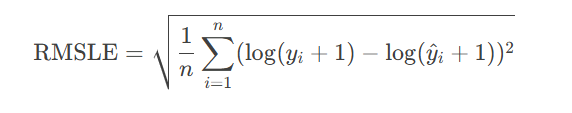
## **Đánh giá hiệu suất mô hình:**

* Để đánh giá độ chính xác của các mô hình dự đoán, bài báo sử dụng chỉ số sai số trung bình bình phương gốc dạng logarit (Root-Mean-Square Logarithmic Error - RMSLE). RMSLE là một thước đo phổ biến trong các bài toán dự đoán giá trị liên tục, đặc biệt là khi giá trị mục tiêu có phân phối lệch (skewed distribution) như giá nhà. Việc sử dụng RMSLE giúp giảm thiểu ảnh hưởng của các giá trị ngoại lai (outliers) và tập trung vào việc đánh giá độ chính xác tương đối của mô hình.

# **Kết quả đạt được và độ đo**

## **Lựa chọn mô hình tối ưu:**

* Sau khi triển khai và so sánh một loạt các mô hình hồi quy khác nhau, bài báo của Yueting Han đã xác định được mô hình Gradient Boosting Regression đã đi đến kết luận về mô hình dự đoán giá nhà tốt nhất. Trong số các mô hình được thử nghiệm (LASSO Regression, Elastic Net Regression, Gradient là mô hình dự đoán giá nhà tốt nhất trong số các mô hình được thử nghiệm. Sự lựa chọn này dựa trên việc đánh giá kỹ lưỡng hiệu suất của Boosting Regression, XGBoost, LightGBM, và Stacked Model), mô hình **Gradient Boosting Regression** nổi lên như là mô hình có hiệu suất vượt trội.
* Mô hình Gradient Boosting Regression đạt được sai số trung bình bình phương gốc các mô hình thông qua các độ đo và phương pháp kiểm định phù hợp.
* **Độ đo đánh giá:** Sai số trung bình bình phương gốc dạng logarit (RMSLE)
* Để đánh giá độ chính xác của các mô hình dự đoán, bài báo sử dụng một độ đo quan trọng là sai số trung bình bình phương gốc dạng logarit (Root-Mean-Square Logarithmic Error – RMSLE). RMSLE được tính toán dựa trên sự khác biệt giữa giá trị dự dạng logarit (RMSLE) thấp nhất, chỉ 0.04615. Điều này cho thấy mô hình có khả năng dự đoán giá nhà với độ chính xác cao, ít bị sai lệch so với giá trị thực đoán (đã được logarit hóa) và giá trị thực tế (cũng đã được logarit hóa). Công thức của RMSLE như sau:



* Trong đó:
* *n* là số lượng mẫu tế.
* *p* giá trị dự đoán (đã được biến đổi logarit).
* *a* là giá trị thực tế (đã được biến đổi logarit).
* **Lựa chọn độ đo: RMSLE** Bài báo sử dụng RMSLE làm thước đo chính để đánh giá hiệu suất của các mô hình dự đoán. Việc lựa chọn RMSLE có những lý do sau:
* Giảm thiểu ảnh hưởng của giá trị ngoại lai trường bất động sản thường có những biến động lớn và các trường hợp đặc biệt.
* **Phù hợp với dữ liệu có phân phối lệch (skewed distribution):** Giá nhà thường có phân phối lệch phải (positive skew), nghĩa là có nhiều nhà giá thấp và trung bình hơn là nhà giá cao. Bi (Outliers):\*\* Trong dữ liệu giá nhà, thường có những ngôi nhà có giá trị rất cao hoặc rất thấp so với phần còn lại. Những giá trị ngoại lai này có thể gây ra sai lệch lớn nếu sử dụng các độ đo như sai số trung bình bình phương (RMSE). RMSLE, bằng cách sử dụng logến đổi logarit giúp làm cho phân phối của giá nhà trở nên gần với phân phối chuẩn hơn, giúp các mô hình hồi quy hoạt động hiệu quả hơn.
* **Đánh giá lỗi tương đối:** RMSLE đo lường sai số tương đối giữa giá trị dự đoán và giá trị thực tế, thay vì sai số tuyệt đối. Điều này có nghĩa là RMSLE sẽ phạt nặng hơn các trường hợp dự đoán sai lệch lớn so với giá trị thực, đặc biệt là đối với những ngôi nhà có giá trị thấp.
* **Kết quả cụ thể của mô hình Gradient Boosting Regression:** Mô hình Gradient Boostingarit của giá trị thực tế và giá trị dự đoán, giúp giảm thiểu ảnh hưởng của các giá trị ngoại lai này.
* **Phù hợp với phân phối lệch (Skewed Distribution):** Giá nhà thường có phân phối lệch phải (positively skewed), tức là có nhiều ngôi nhà giá thấp Regression đã đạt được kết quả RMSLE thấp nhất là 0.04615 trên tập dữ liệu kiểm tra (test set). Điều này cho thấy mô hình có khả năng dự đoán giá nhà với độ chính xác tương đối cao.

## **Kiểm định chéo (Cross-Validation):**

* Để và trung bình ít ngôi nhà giá cao. RMSLE, thông qua phép biến đổi logarit, giúp làm cho phân phối trở nên gần với phân phối chuẩn hơn, từ đó cải thiện hiệu suất của các mô hình hồi quy tuyến tính. RMSLE được tính như sau:
* Tính log đảm bảo tính ổn định và độ tin cậy của mô hình, bài báo còn sử dụng phương pháp kiểm định chéo (cross-validation). Phương pháp này chia tập dữ liệu huấn luyện thành nhiều phần nhỏ (folds), sau đó huấn luyện và đánh giá mô hình trên các tổ hợp khác nhau của các phần nàyarit tự nhiên (natural logarithm) của giá trị dự đoán và giá trị thực tế (cộng thêm 1 để tránh logarit của 0).
* Tính sai số bình phương (squared error) giữa các giá trị logarit.
* Tính trung bình của các sai số bình phương.. Kết quả kiểm định chéo của mô hình Gradient Boosting Regression cho thấy giá trị trung bình của RMSLE là 0.1174. Giá trị này, mặc dù cao hơn so với RMSLE trên tập kiểm tra, nhưng vẫn cho thấy mô hình có khả năng tổng quát hóa tốt, không bị quá khớp (overfitting) với dữ liệu huấn luyện.

| Models/evaluation | rmlse\_cv(mean, std) | rmlse |
| --- | --- | --- |
| LASSO Regression | 0.1189(0.0055) | 0.11294 |
| Elastic Net Regression | 0.1188(0.0055) | 0.11275 |
| Gradient Boosting Regression | 0.1174(0.0067) | 0.04615 |
| XBoost | 0.1164(0.0078) | 0.08093 |
| LightGBM | 0.1192(0.0066) | 0.07568 |
| Stacked Model | 0.1119(0.0074) | 0.66917 |

**Table 1**.Final Evaluations of the Six Models

# **Tính ứng dụng thực tế**

## **Hỗ trợ người mua nhà đưa ra quyết định sáng suốt:**

* **Hiểu rõ các yếu tố quyết định giá trị:** Bài báo giúp người mua tiềm năng "giải mã" các yếu tố phức tạp ảnh hưởng đến giá nhà. Thay vì chỉ dựa vào cảm tính hoặc thông tin chung chung, người mua có thể hiểu rõ hơn về tầm quan trọng của các đặc điểm như chất lượng tổng thể ('OverallQual'), diện tích sinh hoạt ('GrLivArea'), khu vực lân cận ('Neighborhood'), và các yếu tố khác.
* **Đánh giá khách quan hơn:** Với kiến thức này, người mua có thể so sánh các ngôi nhà khác nhau một cách khách quan hơn, xem xét các yếu tố quan trọng một cách có hệ thống, và đưa ra quyết định dựa trên thông tin cụ thể, thay vì chỉ dựa vào giá cả.
* **Thương lượng hiệu quả hơn:** Khi đã hiểu rõ về các yếu tố ảnh hưởng đến giá, người mua có thể tự tin hơn trong quá trình thương lượng, đưa ra các lập luận có cơ sở để đạt được mức giá hợp lý.

## **Cung cấp công cụ hỗ trợ ra quyết định cho nhà đầu tư bất động sản:**

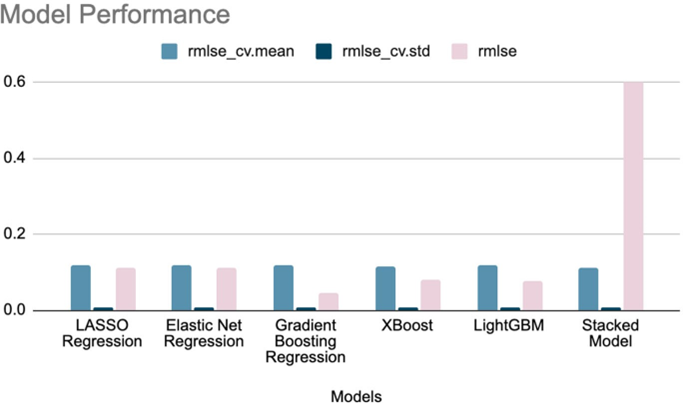
* **Dự đoán tiềm năng sinh lời:** Mô hình dự đoán giá nhà có thể giúp các nhà đầu tư ước tính giá trị tương lai của một bất động sản, từ đó đánh giá tiềm năng sinh lời của khoản đầu tư.
* **Xác định cơ hội đầu tư:** Bằng cách phân tích các đặc điểm của ngôi nhà và so sánh với giá thị trường, nhà đầu tư có thể xác định các bất động sản bị định giá thấp (undervalued) và có tiềm năng tăng giá trong tương lai.
* **Quản lý rủi ro:** Mô hình dự đoán cũng có thể được sử dụng để đánh giá rủi ro của một khoản đầu tư, giúp nhà đầu tư đưa ra quyết định thận trọng hơn.

## **Hỗ trợ hoạch định chính sách và quản lý thị trường:**

* **Dự báo xu hướng thị trường:** Các cơ quan quản lý nhà nước, các tổ chức tài chính, và các công ty nghiên cứu thị trường có thể sử dụng mô hình dự đoán giá nhà để theo dõi và dự báo xu hướng của thị trường bất động sản. Điều này giúp họ có cái nhìn tổng quan và đưa ra các chính sách, quyết định phù hợp.
* **Điều tiết thị trường:** Trong trường hợp thị trường có dấu hiệu bất ổn (ví dụ: giá nhà tăng quá nhanh), các cơ quan quản lý có thể sử dụng thông tin từ mô hình dự đoán để đưa ra các biện pháp can thiệp kịp thời, như điều chỉnh chính sách thuế, lãi suất, hoặc các quy định về xây dựng.
* **Phát triển đô thị:** Thông tin về các yếu tố ảnh hưởng đến giá nhà (ví dụ: vị trí, tiện ích công cộng) có thể được sử dụng để định hướng quy hoạch và phát triển đô thị, đảm bảo sự phát triển bền vững và đáp ứng nhu cầu của người dân.

| Id | SalePrice |
| --- | --- |
| 1461 | 121439.42 |
| 1462 | 160816.69 |
| 1463 | 187468.2 |

**Table 2**.Top rows of the Final Prediction Using Stacked Model



**Figure 5.**.Model Performance

# **Các bài khảo sát liên quan**

## **Rahidi et al. (2015): Phân Loại Các Yếu Tố Ảnh Hưởng**

* Một trong những nghiên cứu quan trọng được đề cập là công trình của Rahidi và các đồng nghiệp, được công bố vào năm 2015 [[2](#_9v2mq7ove4m6)]. Nghiên cứu này tập trung vào việc phân tích các yếu tố ảnh hưởng đến giá nhà ở tại Indonesia. Điểm nổi bật của nghiên cứu này là việc phân loại các yếu tố ảnh hưởng thành ba nhóm chính, cung cấp một khung phân tích có hệ thống:
* **Điều kiện vật lý (Physical Conditions):** Nhóm này bao gồm các đặc điểm hữu hình, có thể quan sát được của ngôi nhà, như diện tích, số phòng ngủ, năm xây dựng, loại vật liệu, v.v.
* **Khái niệm (Concept):** Nhóm này đề cập đến các yếu tố thiết kế, ý tưởng, hoặc phong cách của ngôi nhà, thường được các nhà phát triển sử dụng để thu hút người mua. Ví dụ: phong cách kiến trúc, thiết kế "nhà thông minh", "không gian xanh", v.v.
* **Vị trí (Location):** Nhóm này bao gồm các yếu tố liên quan đến vị trí địa lý của ngôi nhà và môi trường xung quanh, như khu vực lân cận, khoảng cách đến các tiện ích công cộng, chất lượng môi trường, v.v.
* Cách tiếp cận phân loại này của Rahidi et al. [[2](#_9v2mq7ove4m6)] đã cung cấp một nền tảng hữu ích cho Yueting Han trong việc tổ chức và phân tích các đặc điểm của bộ dữ liệu Ames Housing.

## **Zhou (2021): Kết Hợp Hồi Quy Đa Thức và Tối Ưu Hóa**

* Bài báo cũng đề cập đến nghiên cứu của Zhou vào năm 2021 [[3](#_2gxkyhde81k)], trong đó tác giả sử dụng một phương pháp kết hợp giữa hồi quy đa thức (polynomial regression) và tối ưu hóa bầy đàn (particle swarm optimization - PSO) để dự đoán giá nhà.
* **Hồi quy đa thức:** Đây là một dạng mở rộng của hồi quy tuyến tính, cho phép mô hình hóa các mối quan hệ phi tuyến giữa các biến số.
* **Tối ưu hóa bầy đàn (PSO):** Đây là một thuật toán tối ưu hóa dựa trên mô phỏng hành vi của các đàn chim hoặc đàn cá, được sử dụng để tìm ra các tham số tối ưu cho mô hình hồi quy.
* Việc Zhou [[3](#_2gxkyhde81k)] sử dụng PSO cho thấy một hướng tiếp cận khác trong việc giải quyết bài toán dự đoán giá nhà, tập trung vào việc tìm kiếm các mô hình phức tạp hơn để nắm bắt các mối quan hệ phi tuyến trong dữ liệu.

## **Alfiyatin et al. (2017): Một Nghiên Cứu Tương Tự**

* Một nghiên cứu khác có liên quan là công trình của Alfiyatin và các đồng nghiệp vào năm 2017 [[4](#_jt6kt0wdn5ax)]. Nghiên cứu này cũng sử dụng kết hợp phân tích hồi quy và tối ưu hóa bầy đàn (PSO) để dự đoán giá nhà. Việc tham khảo cả nghiên cứu của Zhou [[3](#_2gxkyhde81k)] và Alfiyatin et al. [[4](#_jt6kt0wdn5ax)] cho thấy sự quan tâm của Yueting Han đến các phương pháp tối ưu hóa và khả năng ứng dụng của chúng trong lĩnh vực dự đoán giá nhà.
* Sự khác nhau ở phần định nghĩa ở phần Literature Review (Định nghĩa của Physical conditions) của các tác giả:
* **Trong bài của Han**:Định nghĩa:  
  The category of Physical conditions are features possessed by a property which can be observed by people, such as the number of bedrooms, the area of the living room, and the year when the house is built.
* **Trong bài của Rahidi.et al**:Physical quality consists of  
  variables such as: number of bedrooms, number of bathrooms, land area, building  
  area, availability of electricity (in Watt) and construction quality (wall, floor and  
  roof) of real estate.
* Có nghĩa là phần định nghĩa có sự tương đồng, và định nghĩa của Han ngắn gọn hơn

# **Kết Luận**

## **Thành công và đóng góp:**

* Bài báo "Price Prediction of Ames Housing Through Advanced Regression Techniques" của Yueting Han đã đạt được những thành công đáng kể trong việc áp dụng các kỹ thuật hồi quy và học máy tiên tiến để giải quyết bài toán dự đoán giá nhà tại Ames, Iowa. Bằng cách kết hợp một loạt các phương pháp, từ phân tích tương quan, kỹ thuật đặc trưng, đến các mô hình hồi quy phức tạp như Gradient Boosting Regression, bài báo đã xây dựng được một mô hình dự đoán có độ chính xác cao. Mô hình Gradient Boosting Regression, với sai số trung bình bình phương gốc dạng logarit (RMSLE) thấp nhất, đã chứng minh được khả năng dự đoán giá nhà một cách hiệu quả.
* Không chỉ dừng lại ở việc xây dựng mô hình, bài báo còn làm sáng tỏ các yếu tố quan trọng ảnh hưởng đến giá trị bất động sản. Việc xác định được các đặc điểm như 'OverallQual' (chất lượng tổng thể), 'GrLivArea' (diện tích sinh hoạt), 'YearBuilt' (năm xây dựng), và 'TotalBsmtSF' (diện tích tầng hầm) có tương quan mạnh với giá nhà cung cấp những thông tin hữu ích cho cả người mua, người bán và các nhà đầu tư.

## **Hạn chế và thách thức:**

* Tuy nhiên, như bất kỳ nghiên cứu khoa học nào, bài báo cũng thừa nhận một số hạn chế và thách thức cần được xem xét:
* **Phân loại đặc điểm thủ công (Manual Feature Categorization):** Việc phân loại các đặc điểm thành các nhóm như 'Building', 'Space', và 'Location' được thực hiện thủ công bởi tác giả. Điều này có thể dẫn đến sự chủ quan và không nhất quán, vì các nhà nghiên cứu khác có thể có cách phân loại khác. Hơn nữa, việc phân loại thủ công có thể bỏ sót các mối quan hệ phức tạp giữa các đặc điểm mà các phương pháp tự động có thể phát hiện.
* **Kích thước mẫu tương đối nhỏ (Relatively Small Sample Size):** Bộ dữ liệu Ames Housing, mặc dù được coi là phong phú và đa dạng, nhưng chỉ bao gồm 1460 mẫu trong tập huấn luyện. Điều này có thể giới hạn khả năng tổng quát hóa của mô hình, đặc biệt là khi áp dụng cho các thị trường bất động sản khác có đặc điểm khác biệt.
* **Bỏ qua các yếu tố kinh tế vĩ mô và thị trường (Ignoring Macroeconomic and Market Factors):** Mô hình dự đoán trong bài báo chỉ tập trung vào các đặc điểm của ngôi nhà mà không xem xét đến các yếu tố kinh tế vĩ mô và các biến động của thị trường bất động sản nói chung. Ví dụ, các yếu tố như lãi suất, tỷ lệ lạm phát, chính sách thuế, tình trạng cung cầu của thị trường, và các sự kiện kinh tế - xã hội có thể có ảnh hưởng đáng kể đến giá nhà, nhưng không được đưa vào mô hình.

## **Hướng phát triển tương lai:**

* Để khắc phục những hạn chế này và nâng cao hơn nữa hiệu quả của mô hình dự đoán, bài báo đã đề xuất một số hướng phát triển trong tương lai:
* **Mở rộng quy mô dữ liệu (Expanding Data Scale):** Thu thập thêm dữ liệu từ các nguồn khác nhau, bao gồm cả dữ liệu lịch sử và dữ liệu từ các khu vực khác, có thể giúp cải thiện độ chính xác và khả năng tổng quát hóa của mô hình.
* **Tích hợp các yếu tố kinh tế vĩ mô (Integrating Macroeconomic Factors):** Bổ sung các biến số kinh tế vĩ mô vào mô hình, như lãi suất, tỷ lệ thất nghiệp, GDP, chỉ số giá tiêu dùng, v.v., có thể giúp mô hình nắm bắt được các biến động của thị trường và đưa ra dự đoán chính xác hơn.
* **Áp dụng các kỹ thuật học sâu (Deep Learning):** Các kỹ thuật học sâu (deep learning), như mạng nơ-ron nhân tạo (artificial neural networks), có thể tự động học các đặc trưng phức tạp từ dữ liệu và mô hình hóa các mối quan hệ phi tuyến, có khả năng cải thiện đáng kể độ chính xác của mô hình dự đoán.
* **Tự động hóa**: Xây dựng hệ thống phân loại, xử lý dữ liệu một cách tự động và chính xác

# **Tài Liệu Tham Khảo**

##### [1] D. De Cock, "Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project," Journal of Statistics Education, vol. 19, no. 3, 2011.

##### [2] R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartoto, and I. B. Syamwil, "Factors influencing the price of housing in Indonesia," International Journal of Housing Market Analysis, 2015.

##### [3] C. Zhou, "House price prediction using polynomial regression with particle swarm optimization," Journal of Physics: Conference Series, 2021.

##### [4] A. N. Alfiyatin, H. Taufiq, R. E. Febrita, and W. F. Mahmudy, "Modeling house price prediction using regression analysis and particle swarm optimization," International Journal of Advanced Computer Science and Applications, 2017.

##### 