|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC SÀI GÒN**  **KHOA CÔNG NGHỆ THÔNG TIN** |
|  |
| **LAB 3** |
| **KHẢO SÁT BÀI BÁO CHO DỮ LIỆU AMES HOUSING** |
| Giảng viên hướng dẫn: Ths. Đỗ Như Tài |
| 3122410193 - Nguyễn Phan Tuấn Kiệt (trưởng nhóm) |
| 3122410194 – Nguyễn Thế Kiên |
| 3122410200 – Phạm Văn Kiệt |
|  |
| **Thành phố Hồ Chí Minh, tháng 02/2025** |

**MỤC LỤC**  
**[1. Thông tin bài báo 3](#_4fi0jop3dd94)**

**[2. Nguồn dữ liệu 3](#_ibpmxxvomu1e)**

**[3. Từ điển dữ liệu 4](#_fb2vapbmec3w)**

**[4. Các bài toán liên quan 5](#_le8n9vsuhmo4)**

[1. Dự đoán giá nhà (Bài toán chính): 5](#_pfqgv5qoi1tm)

[2. Phân tích mức độ quan trọng của đặc trưng (Feature Importance Analysis): 6](#_3cpu9thnr4i4)

[3. Phân tích và xử lý dữ liệu thiếu (Missing Data Analysis and Handling): 7](#_t6fbfg8pqy64)

[4. Phân tích và xử lý dữ liệu ngoại lai( Outliers Analysis and Handling): 7](#_lqrzobib6986)

[5. Tạo đặc trưng mới (Feature Engineering): 9](#_ue9k9q9z1qpt)

**[5. Các Khảo Sát Liên Quan 9](#_7tpv4z2p6suj)**

[1. Phân Tích từ Cuộc Thi Kaggle 10](#_t07wvn35l0xn)

[2. Các Nghiên Cứu Cá Nhân trên GitHub 10](#_13vphxad0f2t)

[3. Tài Nguyên Giáo Dục và Các Nguồn Khác 11](#_ke7h3e6xm3zi)

Link paper : <https://www.researchgate.net/publication/382373765_Research_on_predicting_Ames_housing_price_based_on_forward_selection_regression_and_principal_component_regression/fulltext/669a74208dca9f441b884557/Research-on-predicting-Ames-housing-price-based-on-forward-selection-regression-and-principal-component-regression.pdf?origin=publication_detail&_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uRG93bmxvYWQiLCJwcmV2aW91c1BhZ2UiOiJwdWJsaWNhdGlvbiJ9fQ>

# 

# **Thông tin bài báo**

- **Link paper :** <https://www.researchgate.net/publication/382373765_Research_on_predicting_Ames_housing_price_based_on_forward_selection_regression_and_principal_component_regression/fulltext/669a74208dca9f441b884557/Research-on-predicting-Ames-housing-price-based-on-forward-selection-regression-and-principal-component-regression.pdf?origin=publication_detail&_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uRG93bmxvYWQiLCJwcmV2aW91c1BhZ2UiOiJwdWJsaWNhdGlvbiJ9fQ>

**- Tên bài báo:** "Nghiên cứu dự đoán giá nhà ở Ames dựa trên hồi quy lựa chọn tiến (forward selection regression) và hồi quy thành phần chính (principal component regression)"

**- Tác giả:** Ziwen Wang

**- Nơi xuất bản:** Kỷ yếu Hội nghị Quốc tế lần thứ 4 về Xử lý Tín hiệu và Học máy. DOI: 10.54254/2755-2721/45/20241031

- **Link bài báo:** [ResearchGate](https://www.researchgate.net/publication/382373765_Research_on_predicting_Ames_housing_price_based_on_forward_selection_regression_and_principal_component_regression/fulltext/669a74208dca9f441b884557/Research-on-predicting-Ames-housing-price-based-on-forward-selection-regression-and-principal-component-regression.pdf?origin=publication_detail&_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uRG93bmxvYWQiLCJwcmV2aW91c1BhZ2UiOiJwdWJsaWNhdGlvbiJ9fQ)

# **2. Nguồn dữ liệu**

- Dữ liệu được sử dụng trong nghiên cứu này là bộ dữ liệu Ames Housing, được Dean De Cock tổng hợp vào năm 2011. Bộ dữ liệu này chứa thông tin về 2930 giao dịch bất động sản nhà ở riêng lẻ tại Ames, Iowa, Hoa Kỳ, trong khoảng thời gian từ năm 2006 đến năm 2010. Nó được coi là phiên bản cải tiến và cập nhật của bộ dữ liệu nổi tiếng Boston Housing Dataset. Bài báo cho biết nguồn gốc của bộ dữ liệu này là từ nền tảng Kaggle, một cộng đồng trực tuyến dành cho các nhà khoa học dữ liệu và những người đam mê học máy, nơi người dùng có thể tìm, chia sẻ và phân tích các bộ dữ liệu.

- Bộ dữ liệu Ames Housing do Dean De Cock tổng hợp năm 2011, chứa 2930 giao dịch nhà ở tại Ames, Iowa từ 2006 đến 2010. Bộ dữ liệu này được sử dụng rộng rãi trên nền tảng Kaggle.

- Link Kaggle: [House Prices: Advanced Regression Techniques](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview)

# **3. Từ điển dữ liệu**

**-** Bài báo cung cấp một phần từ điển dữ liệu, nhưng không đầy đủ, trong Bảng 1 và phần tiếp theo của bảng đó. Nó chia các đặc trưng (features) thành 4 nhóm chính, và liệt kê một số đặc trưng tiêu biểu cho mỗi nhóm:

**Môi trường (Environment):** Mô tả khu vực xung quanh và vị trí của ngôi nhà. Một số ví dụ:

* MSZoning: Phân loại khu vực chung (ví dụ: khu dân cư mật độ thấp, khu thương mại).
* LotFrontage: Chiều rộng mặt tiền của lô đất (tính bằng feet).
* LotConfig: Vị trí lô đất so với các lô đất khác (ví dụ: lô đất bên trong, lô góc).
* Condition1: Sự gần gũi với các điều kiện khác nhau (ví dụ: gần đường chính, gần công viên).

**Cấu hình (Configuration):** Mô tả bố cục và các hệ thống cơ bản của ngôi nhà. Ví dụ:

* BsmtFinSF1: Diện tích hoàn thiện của tầng hầm loại 1 (tính bằng feet vuông).
* BsmtUnfSF: Diện tích chưa hoàn thiện của tầng hầm (tính bằng feet vuông).
* TotalBsmtSF: Tổng diện tích tầng hầm (tính bằng feet vuông).
* Electrical: Hệ thống điện (ví dụ: tiêu chuẩn, cầu dao).
* BsmtFullBath: Số lượng phòng tắm đầy đủ ở tầng hầm.
* FullBath: Số lượng phòng tắm đầy đủ ở trên mặt đất.
* GarageType: Loại ga ra (ví dụ: gắn liền, tách rời).
* GarageYrBlt: Năm xây dựng ga ra.

- Nguồn tham khảo chi tiết về từ điển dữ liệu có thể tìm thấy tại [Feature documentation by Dean De Cock](https://jse.amstat.org/v19n3/decock/DataDocumentation.txt)

# **4. Các bài toán liên quan**

Trong bài báo, trọng tâm chính là bài toán **dự đoán giá nhà ở Ames, Iowa**, sử dụng kỹ thuật hồi quy. Dựa vào nội dung bài báo và kiến thức chung về khoa học dữ liệu trên thị trường bất động sản, ta có thể suy luận ra nhiều bài toán liên quan khác có thể được giải quyết (hoặc hỗ trợ giải quyết) với bộ dữ liệu Ames Housing, cùng với mô tả chi tiết:

## **1. Dự đoán giá nhà (Bài toán chính):**

**Mô tả:** Đây là bài toán hồi quy (regression), trong đó mục tiêu là ước tính giá trị của biến SalePrice (giá bán nhà) dựa trên các biến độc lập khác (đặc trưng của ngôi nhà và khu vực). Mô hình được huấn luyện trên dữ liệu lịch sử và sau đó được sử dụng để dự đoán giá của những ngôi nhà mới.

**Chi tiết:**

**Đầu vào (Inputs):** Các đặc trưng của ngôi nhà (ví dụ: diện tích, số phòng ngủ, chất lượng, vị trí,...). Đây là các cột dữ liệu từ Bảng 1 (và các cột khác không được liệt kê trong bài báo) trong bộ dữ liệu Ames Housing.

**Đầu ra (Output):** Giá bán dự đoán của ngôi nhà (SalePrice). Đây là một giá trị số liên tục (continuous numerical value).

**Mô hình (Models):** Bài báo tập trung vào hai mô hình: hồi quy lựa chọn tiến (forward selection regression) và hồi quy thành phần chính (principal component regression - PCR). Tuy nhiên, nhiều mô hình hồi quy khác có thể được áp dụng, như: hồi quy tuyến tính (linear regression), cây quyết định (decision trees), rừng ngẫu nhiên (random forests), mạng nơ-ron (neural networks), máy vectơ hỗ trợ (support vector machines - SVM), và các thuật toán boosting (ví dụ: XGBoost, LightGBM).

**Đánh giá (Evaluation):** RMSE (Root Mean Squared Error) và R-squared (hoặc adjusted R-squared) là các độ đo phổ biến để đánh giá hiệu suất của mô hình hồi quy. Ngoài ra, các độ đo như MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error) cũng có thể được sử dụng.

**Nguồn tham khảo:** [Kaggle competition](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview)

## **2. Phân tích mức độ quan trọng của đặc trưng (Feature Importance Analysis):**

**Mô tả:** Xác định những đặc trưng nào có ảnh hưởng lớn nhất đến giá nhà. Điều này giúp hiểu rõ hơn về các yếu tố thúc đẩy giá trị bất động sản và có thể hỗ trợ trong việc ra quyết định đầu tư, định giá, hoặc cải tạo nhà.

**Chi tiết:**

**Đầu vào:** Mô hình hồi quy đã được huấn luyện và bộ dữ liệu.

**Đầu ra:** Danh sách các đặc trưng được sắp xếp theo mức độ quan trọng, thường được biểu diễn bằng biểu đồ hoặc bảng.

**Phương pháp:**

* **Hồi quy tuyến tính:** Hệ số của các biến (coefficients) trong mô hình tuyến tính cho biết mức độ ảnh hưởng của từng biến (sau khi đã chuẩn hóa dữ liệu).
* **Cây quyết định và rừng ngẫu nhiên:** Các thuật toán này cung cấp các độ đo mức độ quan trọng của đặc trưng dựa trên cách chúng được sử dụng để phân chia dữ liệu trong cây.
* **Permutation Importance:** Một kỹ thuật đánh giá mức độ quan trọng của đặc trưng bằng cách hoán vị ngẫu nhiên giá trị của từng đặc trưng và xem xét sự thay đổi trong hiệu suất của mô hình.
* **SHAP values:** Giá trị từ lý thuyết trò chơi, cho biết sự đóng góp của từng đặc trưng đối với một dự đoán cụ thể.

**Nguồn tham khảo:** Các phương pháp như SHAP values ([Eugene Yan's GitHub project](https://github.com/eugeneyan))

## **3. Phân tích và xử lý dữ liệu thiếu (Missing Data Analysis and Handling):**

**Mô tả**: xác định vị trí, kiểu bị thiếu (ngẫu nhiên hoàn toàn (missing completely at random - MCAR), ngẫu nhiên (missing at random - MAR) hay không ngẫu nhiên(not missing at random - NMAR), số lượng giá trị thiếu và đưa ra cách giải quyết

**Chi tiết:**

* Xác định các cột có missing value, bao nhiêu missing value mỗi cột
* Phân tích kiểu dữ liệu bị thiếu, từ đó đưa ra giải pháp: loại bỏ dữ liệu missing nếu số lượng nhỏ và không quá quan trọng, giữ lại và tìm giá trị thay thế nếu quan trọng và có thể lấy được giá trị đó (Ví dụ, các ô NA mang ý nghĩa cụ thể), hoặc các trường hợp không lấy được dữ liệu missing.

**Nguồn tham khảo:** Phương pháp xử lý missing data được đề cập trong [Tidy Modeling with R](https://www.tmwr.org/) và [Kaggle notebooks](https://www.kaggle.com/code/attackgnome/house-prices-advanced-regression-techniques)

## **4. Phân tích và xử lý dữ liệu ngoại lai( Outliers Analysis and Handling):**

**Mô tả:** Xác định các điểm dữ liệu "ngoại lai" (outliers) có giá trị bất thường so với phần còn lại của dữ liệu. Các ngoại lai có thể ảnh hưởng đến hiệu suất của mô hình và cần được xử lý một cách thích hợp.

**Chi tiết:**

**Phát hiện ngoại lai:**

* **Phương pháp trực quan:** Sử dụng biểu đồ hộp (box plots), biểu đồ phân tán (scatter plots) để trực quan hóa dữ liệu và xác định các điểm nằm ngoài phạm vi thông thường.
* **Phương pháp thống kê:** Sử dụng các quy tắc dựa trên khoảng tứ phân vị (interquartile range - IQR), độ lệch chuẩn (standard deviation), hoặc các phương pháp thống kê khác để xác định các điểm dữ liệu bất thường.

**Xử lý ngoại lai:**

* **Loại bỏ:** Xóa các điểm ngoại lai khỏi dữ liệu (cần thận trọng, vì có thể làm mất thông tin).
* **Thay thế:** Thay thế giá trị ngoại lai bằng một giá trị hợp lý hơn (ví dụ: giá trị trung bình, trung vị, hoặc giá trị được dự đoán từ một mô hình khác).
* **Biến đổi dữ liệu:** Áp dụng các phép biến đổi toán học (ví dụ: logarit, căn bậc hai) để giảm thiểu ảnh hưởng của ngoại lai.
* **Giữ lại và dùng model khác:** Đôi khi, ngoại lai mang thông tin có ý nghĩa.

**Nguồn tham khảo:** Các kỹ thuật phát hiện và xử lý ngoại lai được mô tả rõ trong [Kaggle notebooks](https://www.kaggle.com/code/attackgnome/house-prices-advanced-regression-techniques) và tài liệu giáo dục [Tidy Modeling with R](https://www.tmwr.org/)

## 

## **5. Tạo đặc trưng mới (Feature Engineering):**

**Mô tả:** Tạo ra các đặc trưng mới từ các đặc trưng hiện có để cải thiện hiệu suất của mô hình. Việc này có thể bao gồm kết hợp các đặc trưng, tạo các biến tương tác, hoặc trích xuất thông tin từ các biến dạng văn bản hoặc thời gian.

**Chi tiết:**

**Kết hợp các đặc trưng:** Ví dụ, tạo một biến mới là tổng diện tích của tất cả các tầng, hoặc tỷ lệ giữa diện tích sử dụng và diện tích lô đất.

**Tạo biến tương tác:** Ví dụ, tạo một biến mới là tích của OverallQual và GrLivArea để thể hiện mối quan hệ tương tác giữa chất lượng tổng thể và diện tích sinh hoạt.

**Trích xuất thông tin:** Ví dụ, từ biến YrSold (năm bán), có thể tạo các biến mới như "tuổi của ngôi nhà khi bán" (năm bán trừ năm xây dựng).

**One-hot encoding:** Các features ở dạng categorical thì chuyển về dạng số

**Nguồn tham khảo:** Các kỹ thuật feature engineering chi tiết được trình bày trong các notebook Kaggle ([House Prices: Advanced Regression Techniques](https://www.kaggle.com/code/attackgnome/house-prices-advanced-regression-techniques))[5](https://www.kaggle.com/code/attackgnome/house-prices-advanced-regression-techniques) và tài liệu giáo dục như [Scikit-learn MOOC của Inria](https://www.inria.fr/en/mooc-scikit-learn)

# **5. Các Khảo Sát Liên Quan**

Bộ dữ liệu Ames Housing đã được nghiên cứu và phân tích rộng rãi trong nhiều bối cảnh khác nhau, từ các cuộc thi học máy, dự án cá nhân, đến các tài nguyên giáo dục. Phần này sẽ tổng hợp các khảo sát và phân tích đáng chú ý từ các nguồn đã được cung cấp, bao gồm cuộc thi Kaggle, các dự án trên GitHub, và các tài liệu học thuật.

## 

## **1. Phân Tích từ Cuộc Thi Kaggle**

Cuộc thi **"House Prices: Advanced Regression Techniques"** trên Kaggle ([link](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques)) là một nguồn tài nguyên phong phú với nhiều phân tích từ cộng đồng. Các notebook của người tham gia thường thực hiện phân tích khám phá dữ liệu (EDA) và thử nghiệm các mô hình dự đoán. Trong đó bao gồm:

* Biến mục tiêu **SalePrice** có phân phối lệch phải (right-skewed), thường được biến đổi (ví dụ: log transformation) để phù hợp hơn với các mô hình.
* Các đặc trưng như **OverallQual** (chất lượng tổng thể), **GrLivArea** (diện tích sinh hoạt trên mặt đất), và **Neighborhood** (khu vực lân cận) cho thấy mối tương quan mạnh với giá bán.
* Dữ liệu có giá trị thiếu ở một số cột như **LotFrontage** hay **GarageYrBlt**, đòi hỏi kỹ thuật xử lý phù hợp.
* Các mô hình tiên tiến như **Gradient Boosting** (ví dụ: XGBoost, LightGBM) thường đạt hiệu suất cao hơn so với hồi quy tuyến tính truyền thống.

Những phân tích này không chỉ tập trung vào dự đoán mà còn khám phá sâu các đặc điểm của dữ liệu, cung cấp cái nhìn toàn diện về bộ dữ liệu.

## **2. Các Nghiên Cứu Cá Nhân trên GitHub**

Hai dự án cá nhân trên GitHub mang lại những góc nhìn chi tiết và độc đáo về bộ dữ liệu Ames Housing:

* **Dự án của Melinda Leung** ([link](https://github.com/melindaleung/Ames-Iowa-Housing-Dataset)): Dự án này bao gồm một phân tích EDA toàn diện và các bước kỹ thuật đặc trưng (feature engineering). Tác giả nhấn mạnh việc tạo ra các **tương tác đặc trưng** (interaction terms) như giữa tuổi nhà và tình trạng cải tạo, giúp cải thiện hiệu suất mô hình.
* **Dự án của Eugene Yan** ([link](https://github.com/eugeneyan84/Ames-Housing-Data-Analysis)): Nghiên cứu này tập trung vào khả năng diễn giải mô hình, sử dụng **SHAP values** để đánh giá mức độ ảnh hưởng của từng đặc trưng. Kết quả cho thấy **OverallQual** thường là yếu tố quan trọng nhất, nhưng tác động của nó thay đổi tùy thuộc vào các đặc trưng khác.

Cả hai dự án đều mở rộng phân tích vượt ra ngoài phạm vi cuộc thi Kaggle, cung cấp những cách tiếp cận sáng tạo và chuyên sâu.

## **3. Tài Nguyên Giáo Dục và Các Nguồn Khác**

Bộ dữ liệu Ames Housing cũng xuất hiện trong nhiều tài liệu giáo dục, giúp minh họa các kỹ thuật phân tích dữ liệu và học máy:

* **"Tidy Modeling with R"** ([link](https://www.tmwr.org/ames)): Cuốn sách này sử dụng bộ dữ liệu để hướng dẫn cách tiền xử lý dữ liệu, chọn đặc trưng, và xây dựng mô hình trong R. Nó nhấn mạnh tầm quan trọng của việc xử lý biến phân loại và kiểm tra chéo (cross-validation).
* **Scikit-learn MOOC** ([link](https://inria.github.io/scikit-learn-mooc/python_scripts/datasets_ames_housing.html)): Tài liệu này áp dụng bộ dữ liệu vào việc giảng dạy các công cụ của scikit-learn, tập trung vào quy trình xây dựng mô hình trong Python.
* **Minitab Dataset** ([link](https://support.minitab.com/en-us/datasets/predictive-analytics-data-sets/ames-housing-data/)): Đây là một mô tả cơ bản của bộ dữ liệu kèm theo một số ví dụ phân tích đơn giản, phù hợp cho người mới bắt đầu.

Ngoài ra, một phiên bản khác của bộ dữ liệu trên Kaggle ([link](https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset)) cũng được đề cập, nhưng chủ yếu là nguồn dữ liệu thô với ít phân tích đi kèm.