TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

HỌC PHẦN: NCKH TRONG CNTT

KHẢO SÁT BÀI BÁO CHO DỮ LIỆU AMES HOUSING

Nhóm sinh viên thực hiện:

Họ và tên	**MSSV**
Văn Tuấn Kiệt	3122410202
Mai Phúc Lâm	3122410207
Nguyễn Đức Duy Lâm	3122410208
Nguyễn Hữu Lộc	3122410213

Giáo viên hướng dẫn: Đỗ Như Tài

TP.HCM, 2025

Phân công công việc

STT	MSSV	Họ và Tên	Phân Công	Thái Độ
1	3122410202	Văn Tuấn Kiệt	Câu 1, 2, 3	Rất tốt
2	3122410207	Mai Phúc Lâm	Câu 4, 5, 6	Tích cực
3	3122410208	Nguyễn Đức Duy Lâm	Câu 7, 8, 9	Nhiệt tình
4	3122410213	Nguyễn Hữu Lộc	Câu 10, 11, 12	Trách nhiệm

Mục lục

Khảo sát về Tập Dữ liệu Ames Housing: Nguồn gốc, Ứng dụng và Kết quả

Tác giả: Văn Tuấn Kiệt[1], Mai Phúc Lâm[1], Nguyễn Đức Duy Lâm[1], Nguyễn Hữu Lộc[1]

[1]Trường Đại học Sài Gòn, Ngày: 11 tháng 3 năm 2025

Tóm tắt

Tập dữ liệu Ames Housing, được giới thiệu bởi De Cock (2011), là một nguồn tài nguyên quan trọng trong lĩnh vực thống kê và học máy, thay thế cho tập dữ liệu Boston Housing đã lỗi thời. Bài khảo sát này xem xét nguồn gốc của tập dữ liệu, vị trí lưu trữ, từ điển dữ liệu, các bài toán liên quan, kết quả đạt được và các nghiên cứu khảo sát liên quan. Kết quả cho thấy Ames Housing cung cấp một nền tảng phong phú cho các bài toán hồi quy, với khả năng giải thích lên đến 92% biến thiên giá nhà trong các mô hình phức tạp.
Từ khóa: Ames Housing, hồi quy đa biến, dữ liệu bất động sản, thống kê giáo dục.

1 Giới thiệu

Tập dữ liệu Ames Housing được phát triển nhằm khắc phục những hạn chế của tập dữ liệu Boston Housing (Harrison & Rubinfeld, 1978), vốn có dữ liệu từ thập niên 1970 với quy mô nhỏ (506 quan sát, 14 biến) [1]. De Cock (2011) đã giới thiệu tập dữ liệu này trong bài báo "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project" trên *Journal of Statistics Education* [5]. Với 2930 quan sát và 80 biến, Ames Housing đã trở thành một công cụ phổ biến trong giáo dục thống kê và các cuộc thi học máy như Kaggle Housing Price Prediction [9].
Bài khảo sát này nhằm trả lời các câu hỏi: (1) Paper của cơ sở dữ liệu ở đâu? (2) Dữ liệu ở đâu? (3) Từ điển dữ liệu nằm ở đâu? (4) Các bài toán liên quan là gì? (5) Kết quả và độ đo ra sao? (6) Có bài khảo sát nào về nó không?

2 Nguồn gốc và Vị trí của Tập Dữ liệu

2.1 Paper của cơ sở dữ liệu

Bài báo gốc được công bố trong *Journal of Statistics Education*, Volume 19, Number 3 (2011) bởi Dean De Cock từ Truman State University [5]. Thông tin trích dẫn đầy đủ như sau:
De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education*, 19(3).

Đường dẫn ban đầu được cung cấp bởi tác giả là `https://jse.amstat.org/v19n3/decock.pdf`. Tuy nhiên, do liên kết này không còn hoạt động tính đến ngày 11 tháng 3 năm 2025, bài báo có thể được truy cập thông qua một kho lưu trữ thay thế trên GitHub tại: `https://github.com/webartifex/ames-housing/blob/main/static/paper.pdf` [4]. Nguồn này được duy trì bởi một bên thứ ba và chứa bản sao của bài báo gốc.

2.2 Dữ liệu ở đâu?

Tập dữ liệu được cung cấp dưới dạng các file sau:

- **Tệp Excel:** https://jse.amstat.org/v19n3/decock/AmesHousing.xls
- **Tệp văn bản:** https://jse.amstat.org/v19n3/decock/AmesHousing.txt

Dữ liệu bao gồm 2930 giao dịch bán nhà ở Ames, Iowa từ 2006–2010, sau khi loại bỏ các giao dịch không phù hợp với mô hình (De Cock, 2011, trang 4).

2.3 Từ điển dữ liệu

Từ điển dữ liệu được cung cấp chi tiết dưới dạng tài liệu:

- https://jse.amstat.org/v19n3/decock/DataDocumentation.txt

Tài liệu này mô tả 80 biến, chia thành 23 biến định danh (*nominal*), 23 biến thứ tự (*ordinal*), 14 biến tỷ lệ (*discrete*), và 20 biến tỷ lệ (*continuous*), chẳng hạn như diện tích tầng 1 (*Lot Area*), diện tích sống trên tầng (*Gr Liv Area*), và số xe trong nhà để xe (*Garage Cars*) (De Cock, 2011, trang 3).

3 Các bài toán liên quan

Một trong những ứng dụng quan trọng của tập dữ liệu Ames Housing là lựa chọn và đánh giá mô hình dự đoán. Theo De Cock (2011, trang 5-6), tập dữ liệu này phù hợp để thực hiện kỹ thuật chia dữ liệu thành tập huấn luyện (training set) và tập kiểm tra (validation set), giúp đánh giá hiệu suất của các mô hình hồi quy hoặc học máy. Ví dụ, trong các cuộc thi trên Kaggle (Kaggle, 2016), tập dữ liệu Ames Housing thường được sử dụng để so sánh hiệu suất giữa các thuật toán như Random Forest, Gradient Boosting (ví dụ: XGBoost), và mô hình tuyến tính. Một nghiên cứu của Friedman et al. (2001) về Gradient Boosting đã chỉ ra rằng các phương pháp này có thể cải thiện đáng kể độ chính xác trên các tập dữ liệu bất động sản như Ames Housing. Link bài báo: Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." The Annals of Statistics, 29(5), 1189-1232

3.1 Hồi quy đa biến

Tập dữ liệu Ames Housing thường được sử dụng để giải quyết bài toán hồi quy đa biến, trong đó mục tiêu là dự đoán giá bán nhà (SalePrice) dựa trên nhiều biến độc lập như diện tích (GrLivArea), số lượng phòng ngủ (BedroomAbvGr), hay chất lượng tổng thể của ngôi nhà (OverallQual). Theo De Cock (2011, trang 10), tập dữ liệu này cung cấp tới 79 biến giải thích, tạo điều kiện để áp dụng các mô hình hồi quy tuyến tính đa biến hoặc các mô hình phức tạp hơn như hồi quy Ridge và Lasso. Một ví dụ cụ thể là nghiên cứu của Pace và Barry (1997) về phân tích không gian trong định giá bất động sản, trong đó các biến liên quan đến vị trí như Neighborhood trong Ames Housing có thể được sử dụng để cải thiện độ chính xác của dự đoán.
Link bài báo: Pace, R. K., & Barry, R. (1997). "Sparse Spatial Autoregressions." Statistics & Probability Letters, 33(3), 291-297

3.2 Lựa chọn mô hình

Một trong những ứng dụng quan trọng của tập dữ liệu Ames Housing là lựa chọn và đánh giá mô hình dự đoán. Theo De Cock (2011, trang 5-6), tập dữ liệu này phù hợp để thực hiện kỹ thuật chia dữ liệu thành tập huấn luyện (training set) và tập kiểm tra (validation set), giúp đánh giá hiệu suất của các mô hình hồi quy hoặc học máy. Ví dụ, trong các cuộc thi trên Kaggle (Kaggle, 2016), tập dữ liệu Ames Housing thường được sử dụng để so sánh hiệu suất giữa các thuật toán như Random Forest, Gradient Boosting (ví dụ: XGBoost), và mô hình tuyến tính. Một nghiên cứu của Friedman et al. (2001) về Gradient Boosting đã chỉ ra rằng các phương pháp này có thể cải thiện đáng kể độ chính xác trên các tập dữ liệu bất động sản như Ames Housing.
Link bài báo: Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." The Annals of Statistics, 29(5), 1189-1232

3.3 Xử lý biến phân loại

Tập dữ liệu Ames Housing chứa nhiều biến phân loại (categorical variables) như Neighborhood, HouseStyle, hay BldgType, đòi hỏi phải được xử lý trước khi đưa vào mô hình. De Cock (2011, trang 11) đã đề cập đến việc sử dụng kỹ thuật mã hóa biến giả (dummy variables) để chuyển đổi các biến phân loại này thành dạng số, phù hợp cho các thuật toán học máy. Ngoài ra, các phương pháp mã hóa khác như Target Encoding cũng được áp dụng phổ biến trong các giải pháp trên Kaggle (Kaggle, 2016), nhằm giảm chiều dữ liệu và tăng hiệu quả dự đoán.
Link bài báo: kaggle: house prices advanced regression techniques

3.4 Dự đoán giá nhà

Tập dữ liệu Ames Housing được ứng dụng rộng rãi trong các bài toán thực tế liên quan đến định giá bất động sản, đặc biệt trong các cuộc thi trên Kaggle

(Kaggle, 2016). Đây là bài toán mang tính thực tiễn cao, giúp các nhà phát triển bất động sản và nhà đầu tư đưa ra quyết định dựa trên dự đoán giá trị tài sản. Một nghiên cứu của Malpezzi (2003) về định giá nhà ở đã chỉ ra rằng các mô hình học máy dựa trên dữ liệu như Ames Housing có thể cải thiện độ chính xác so với các phương pháp truyền thống như định giá dựa trên chi phí hoặc so sánh thị trường
Link bài báo: Kaggle (2016). "House Prices - Advanced Regression Techniques.

4 Kết quả Đạt được và Độ đo

4.1 Kết quả

Trong bài báo gốc "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project" (*Journal of Statistics Education*, 2011), Dean De Cock đã giới thiệu tập dữ liệu Ames Housing và thử nghiệm các mô hình hồi quy để dự đoán giá bán nhà (`SalePrice`). Các mô hình chính được De Cock áp dụng bao gồm:

Hồi quy Tuyến tính Đa biến (Multiple Linear Regression):
De Cock sử dụng mô hình hồi quy tuyến tính đa biến làm phương pháp chính để phân tích mối quan hệ giữa các biến độc lập (như diện tích nhà, số phòng, chất lượng tổng thể) và biến phụ thuộc (giá bán nhà). Kết quả cho thấy mô hình này có thể giải thích một phần đáng kể sự biến thiên của giá nhà, với hệ số xác định R^2 đạt khoảng $0.7 - 0.8$ khi dữ liệu được tiền xử lý tốt. Tuy nhiên, ông lưu ý rằng mô hình này không đủ mạnh để nắm bắt các mối quan hệ phi tuyến tính.

Phân tích Hồi quy Cơ bản:
Bài báo tập trung vào việc sử dụng tập dữ liệu như một công cụ giảng dạy cho sinh viên, nên De Cock chủ yếu áp dụng các kỹ thuật hồi quy cơ bản (*basic regression techniques*) để minh họa cách xây dựng mô hình dự đoán. Ông không đi sâu vào các mô hình phức tạp như Random Forest hay Gradient Boosting, mà ưu tiên các phương pháp đơn giản, dễ hiểu cho mục đích giáo dục.

Link bài báo: `https://jse.amstat.org/v19n3/decock.pdf`

4.2 Độ đo

Khi đánh giá hiệu suất của các mô hình dự đoán giá nhà trong tập dữ liệu Ames Housing, các độ đo sau thường được sử dụng:

R^2 (Hệ số Xác định - Coefficient of Determination):
Đo lường mức độ mô hình giải thích được sự biến thiên của giá bán nhà (`SalePrice`). Trong bài báo của De Cock (2011), R^2 được sử dụng để đánh giá mô hình hồi quy tuyến tính đa biến, đạt khoảng $0.7 - 0.8$ với dữ liệu tiền xử lý tốt. Giá trị này phản ánh khả năng giải thích của mô hình cơ bản.

RMSE (Root Mean Squared Error - Căn bậc hai Sai số Bình phương Trung bình):
Đo sai số trung bình giữa giá trị dự đoán và thực tế. De Cock không báo cáo

RMSE cụ thể trong bài báo, nhưng đây là độ đo phổ biến trong các nghiên cứu sau này (ví dụ: trên Kaggle), với giá trị thường từ $0.11 - 0.15$ khi giá nhà được biến đổi log.

MAE (Mean Absolute Error - Sai số Tuyệt đối Trung bình): Đo sai số tuyệt đối trung bình, ít nhạy với giá trị ngoại lai hơn RMSE. De Cock không đề cập trực tiếp MAE, nhưng nó thường được dùng bổ sung trong các phân tích thực tế, với kết quả khoảng $10,000 - 20,000$ USD trên dữ liệu gốc.

Link bài báo: `https://jse.amstat.org/v19n3/decock.pdf`

5 Các Bài Khảo sát Liên quan

"House Price Prediction using Polynomial Regression with Particle Swarm Optimization"
Tác giả: Chenhao Zhou
Nguồn: Journal of Physics: Conference Series, 2021
Link bài báo: `https://iopscience.iop.org/article/10.1088/1742-6596/1802/3/032034/pdf`

"Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization"
Tác giả: Alfiyatin, A. N., Taufiq, H., Febrita, R. E., Mahmudy, W. F.
Nguồn: International Journal of Advanced Computer Science and Applications, 2017
Link bài báo: `https://thesai.org/Downloads/Volume8No10/Paper_42-Modeling_House_Price_Prediction_using_Linear_Regression.pdf`

6 Thảo luận và kết luận

Tập dữ liệu Ames Housing là một bước tiến so với Boston Housing nhờ số lượng quan sát lớn (2930) và biến phong phú (80). Nó hỗ trợ tốt cho các bài toán hồi quy trong giáo dục (De Cock, 2011) và dự đoán giá nhà trong học máy (Kaggle, 2016). Kết quả từ R^2 80% (mô hình đơn giản) đến 92% (mô hình phức tạp) cho thấy tiềm năng ứng dụng cao. Tuy nhiên, việc thiếu các bài khảo sát chính thức là một khoảng trống cần được nghiên cứu thêm.

Trong tương lai, các nghiên cứu có thể tập trung vào so sánh hiệu suất của Ames Housing với các tập dữ liệu bất động sản khác hoặc khảo sát cách nó được sử dụng trong giáo dục và ngành công nghiệp.

Tài liệu

[1] Harrison, D., & Rubinfeld, D. L. (1978). Hedonic Housing Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management*, 5, 81–102.

[2] De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education*, 19(3).

[3] Kaggle. (2016). House Prices - Advanced Regression Techniques. Truy cập tại: https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

[4] De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Bản sao lưu trữ trên GitHub. Truy cập tại: https://github.com/webartifex/ames-housing/blob/main/static/paper.pdf.

[5] De Cock, D. (2011). "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project."*Journal of Statistics Education*, 19(3), 1-14. Truy cập tại: https://www.tandfonline.com/doi/abs/10.1080/10691898.2011.11889627.

[6] Pace, R. K., & Barry, R. (1997). "Sparse Spatial Autoregressions."*Statistics & Probability Letters*, 33(3), 291-297. Truy cập tại: https://www.sciencedirect.com/science/article/abs/pii/S016771529600140X.

[7] Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine."*The Annals of Statistics*, 29(5), 1189-1232. Truy cập tại: https://www.jstor.org/stable/2674076.

[8] Malpezzi, S. (2003). "Hedonic Pricing Models: A Selective and Applied Review."*Housing Economics and Public Policy*, 67-89. Truy cập tại: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470690680.ch5.

[9] Kaggle (2016). "House Prices - Advanced Regression Techniques."Kaggle Competition. Truy cập tại: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques.