

TÓM TẮT BÀI BÁO

Nguồn: Purushotham K, Bangarappa, Ashwini Kodipalli, Trupthi Rao. "Real-Time House Price Predictions with Regression Analysis." IEEE Xplore, 2024. DOI: 979-8-3503-8168-9/24.

1.1 Giới thiệu:

Bài báo trình bày một dự án phát triển website dự đoán giá nhà ở Bangalore, Ấn Độ, nhằm đáp ứng nhu cầu ngày càng tăng về định giá bất động sản chính xác. Bất động sản là một ngành phức tạp với nhiều yếu tố ảnh hưởng đến giá nhà như vị trí, diện tích, số phòng ngủ, phòng tắm, tiện ích và xu hướng thị trường. Việc định giá đúng giúp người mua, người bán và chuyên gia bất động sản đưa ra quyết định tài chính sáng suốt. Dự án tập trung vào việc sử dụng các mô hình máy học, đặc biệt là hồi quy, để dự đoán giá nhà với độ chính xác cao. Website được thiết kế với giao diện thân thiện, minh bạch về cách tạo ra dự đoán, giúp người dùng dễ dàng nhập thông tin và nhận ước tính giá dựa trên phân tích dữ liệu lịch sử và thị trường hiện tại.

1.2 Mục tiêu:

- + Xây dựng mô hình máy học dự đoán giá nhà chính xác, đáng tin cậy dựa trên các đặc trưng như vị trí, diện tích, số phòng ngủ, phòng tắm và xu hướng thị trường.
- + Phát triển website dễ sử dụng, cung cấp dự đoán giá nhanh chóng và minh bạch về các yếu tố ảnh hưởng đến giá trị bất động sản.
- + Hỗ trợ người dùng (người mua, người bán, chuyên gia) điều hướng thị trường bất động sản phức tạp với sự tự tin.

1.3 Dữ liệu:

Nguồn: Dữ liệu được thu thập từ Kaggle.com: "Bengaluru House Price Data".

Đặc trưng:

Thuộc tính	Ý nghĩa
<i>area_type</i>	Loại khu vực (Super built-up, Built-up...)
<i>availability</i>	Tình trạng sẵn có (Ready To Move, 2025...)

<i>location</i>	Khu vực/quận tại Bangalore
<i>size</i>	Số phòng (“2 BHK”)
<i>society</i>	Tên khu dân cư
<i>total_sqft</i>	Diện tích (ft ²), đôi khi là khoảng (“2100-2850”)
<i>bath</i>	Số phòng tắm
<i>balcony</i>	Số ban-công
<i>price</i>	Giá (lakh INR) – biến mục tiêu

1.4 Phương pháp:

- **Tiền xử lý dữ liệu:**

- + Loại bỏ đặc trưng: Các cột không cần thiết như *area_type*, *society*, *balcony*, *availability* bị xóa để tăng hiệu quả mô hình.
- + Xử lý giá trị thiếu: Điền giá trị trung vị cho các cột số hoặc loại bỏ hàng có giá trị thiếu ở cột quan trọng như *location*, *price*.
- + Kỹ thuật đặc trưng: Tạo đặc trưng mới *price_per_sqft* (giá trên mỗi feet vuông) để chuẩn hóa giá nhà.
- + Loại bỏ ngoại lệ: Sử dụng phương pháp trung bình ± 3 độ lệch chuẩn để loại bỏ các điểm dữ liệu bất thường ở các cột như *price*, *total_sqft*.
- + Mã hóa one-hot: Chuyển đổi cột *location* thành định dạng nhị phân để phù hợp với mô hình máy học.
- + Giảm chiều: Giới hạn số lượng vị trí để giảm số lượng đặc trưng, tránh quá khớp.
- + Chia dữ liệu: 80% dùng để huấn luyện, 20% dùng để kiểm tra.

- **Mô hình máy học:**

- + **Linear Regression:** Mô hình hồi quy tuyến tính để dự đoán giá nhà dựa trên mối quan hệ tuyến tính giữa các đặc trưng và giá.
- + **Lasso Regression:** Hồi quy với chính quy hóa L1, giúp loại bỏ các đặc trưng không quan trọng bằng cách đưa hệ số về 0, giảm nguy cơ quá khớp.
- + **Decision Tree:** Mô hình cây quyết định dự đoán giá nhà thông qua các quyết định phân nhánh dựa trên đặc trưng.

- Tối ưu hóa: Sử dụng GridSearchCV để tinh chỉnh siêu tham số (ví dụ: alpha cho Lasso, max_depth và min_samples_split cho Decision Tree) và áp dụng k-fold cross-validation để đảm bảo đánh giá công bằng.
- Công cụ: Sử dụng Python với các thư viện như pandas, scikit-learn để xử lý dữ liệu và xây dựng mô hình.

1.5 Kết quả:

- Hiệu suất mô hình được đánh giá bằng R^2 score (độ phù hợp của mô hình với dữ liệu).
- Bảng hiệu suất mô hình (theo bài báo):
 - Linear Regression: $R^2 = 0.856654$ (cao nhất).
 - Decision Tree: $R^2 = 0.749781$.
 - Lasso Regression: $R^2 = 0.710544$ (thấp nhất).
- Kết quả cho thấy Linear Regression vượt trội trong việc dự đoán giá nhà, phù hợp với đặc điểm của dataset và các đặc trưng đã chọn.

1.6 Kết luận:

Dự án đã thành công trong việc áp dụng các kỹ thuật máy học để dự đoán giá nhà ở Bangalore. Linear Regression được xác định là mô hình hiệu quả nhất, phù hợp cho các ứng dụng thực tế. Website cung cấp một công cụ thân thiện, minh bạch, giúp người dùng tiếp cận thông tin giá cả đáng tin cậy. Dự án không chỉ hỗ trợ quyết định tài chính mà còn tiết kiệm thời gian, chi phí so với việc thuê quản lý bất động sản. Các bước tiền xử lý dữ liệu và kỹ thuật đặc trưng đóng vai trò quan trọng trong việc cải thiện độ chính xác của mô hình.

1.7 Ý nghĩa:

- + **Ứng dụng thực tiễn:** Website giúp người dùng dễ dàng ước tính giá nhà mà không cần kiến thức chuyên sâu về bất động sản.
- + **Tính minh bạch:** Cung cấp thông tin rõ ràng về các yếu tố ảnh hưởng đến giá, xây dựng niềm tin cho người dùng.
- + **Khả năng mở rộng:** Có thể áp dụng cho các thị trường bất động sản khác với dữ liệu phù hợp.

1.8 Mã nguồn

[] Purushotham, K., Bangarappa, Kodipalli, A., & Rao, T. (2024). Real-Time House Price Predictions with Regression Analysis. In 2024 International Conference on

Advances in Data Science, Communication and Computing (ICADCC) (pp. 1-6). IEEE.
DOI: 10.1109/ICADCC59229.2024.10689962