

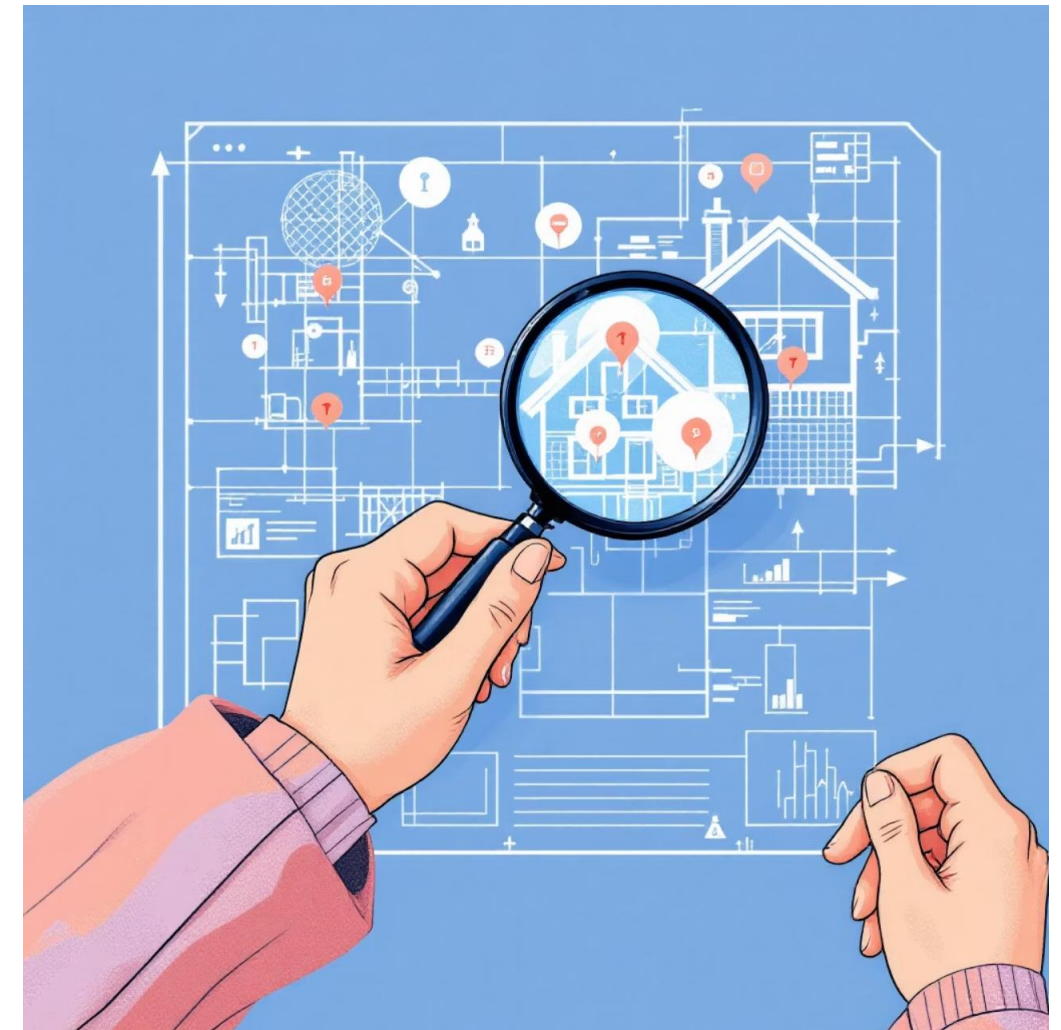


Hệ thống Dự đoán Giá Nhà Dựa trên Dữ liệu Giao dịch Lịch sử

Giới thiệu & Động lực

Việc định giá bất động sản là một thách thức phức tạp, thường phụ thuộc vào đánh giá thủ công dễ bị sai sót. Nhu cầu về các phương pháp tiếp cận dựa trên dữ liệu, minh bạch và hiệu quả, ngày càng trở nên cấp thiết.

- Định giá truyền thống: Chủ quan, tốn thời gian, không nhất quán.
- Phương pháp dựa trên dữ liệu: Khách quan, hiệu quả, có thể mở rộng.
- Mục tiêu: Phát triển hệ thống dự đoán giá nhà chính xác và đáng tin cậy.



Tổng quan về Phương pháp luận

Quy trình từ đầu đến cuối của chúng tôi bao gồm nhiều giai đoạn, đảm bảo thu thập dữ liệu kỹ lưỡng, xử lý hiệu quả và mô hình hóa tối ưu.

1

Thu thập dữ liệu

Tập hợp thông tin từ nhiều nguồn.

2

Tiền xử lý

Làm sạch, chuyển đổi và tạo ra các đặc trưng.

3

Lựa chọn mô hình

Chọn các thuật toán học máy phù hợp.

4

Tối ưu hóa

Điều chỉnh siêu tham số để đạt hiệu suất tốt nhất.

5

Đánh giá

Đánh giá độ chính xác và độ mạnh của mô hình.

Thu thập Dữ liệu: [Nhatot.com](https://nhatot.com)

Chúng tôi đã sử dụng công cụ 'Crawl4AI' để thu thập dữ liệu từ nhatot.com, một trong những nền tảng bất động sản lớn nhất Việt Nam. Dữ liệu bao gồm 7.669 danh sách giao dịch, cung cấp một tập dữ liệu phong phú để phân tích.

- Nguồn: Nhatot.com
- Công cụ: Crawl4AI
- Số lượng danh sách: 7.669
- Chiến lược: Thu thập dữ liệu thời gian thực để phản ánh thị trường biến động.



Phân tích Dữ liệu Khám phá (EDA)

Phân tích ban đầu cho thấy sự phân bố đặc trưng của giá cả và diện tích, cùng với các vấn đề về dữ liệu thiếu.



Phân bố Dài

Giá cả và diện tích cho thấy phân bố dài, với một số lượng nhỏ các giá trị cực đoan.



Giá trị ngoại lai

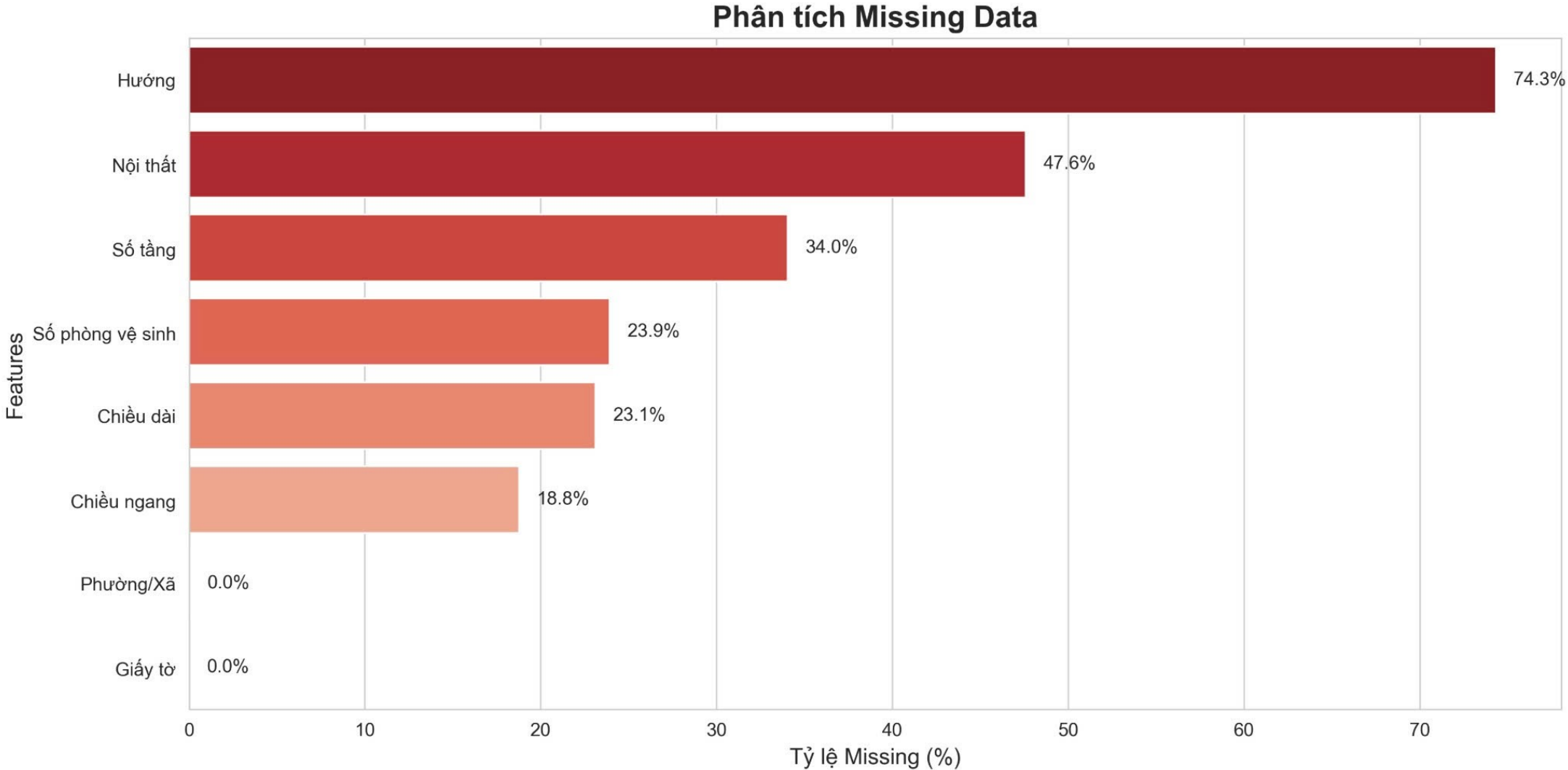
Sự hiện diện của các giá trị ngoại lai đáng kể cần được xử lý cẩn thận trong quá trình tiền xử lý.



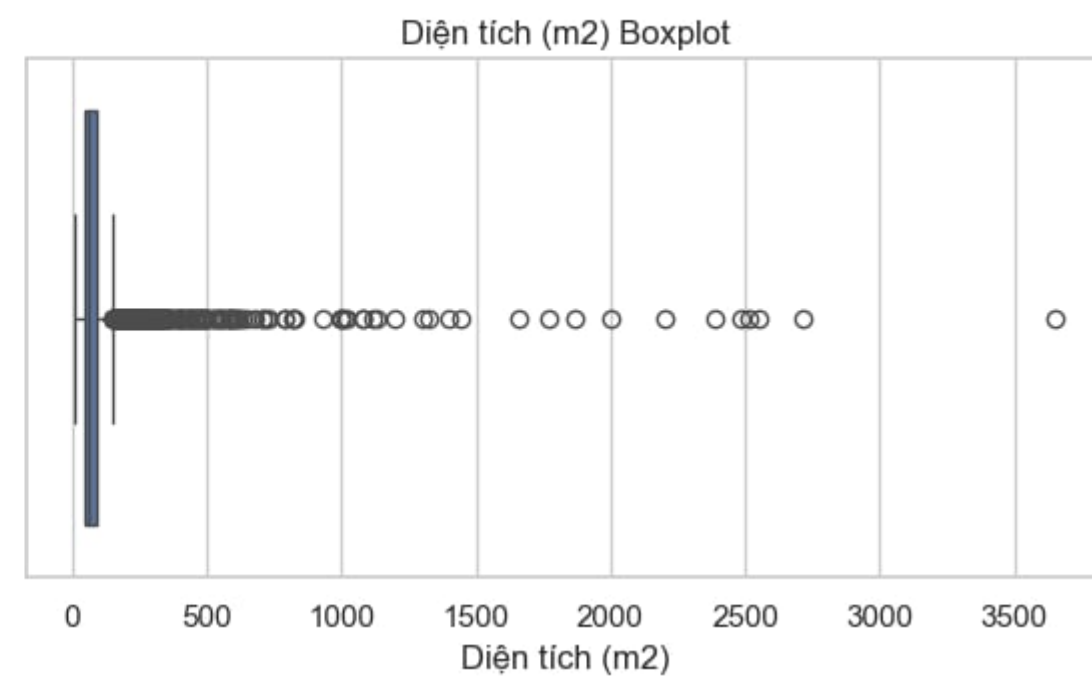
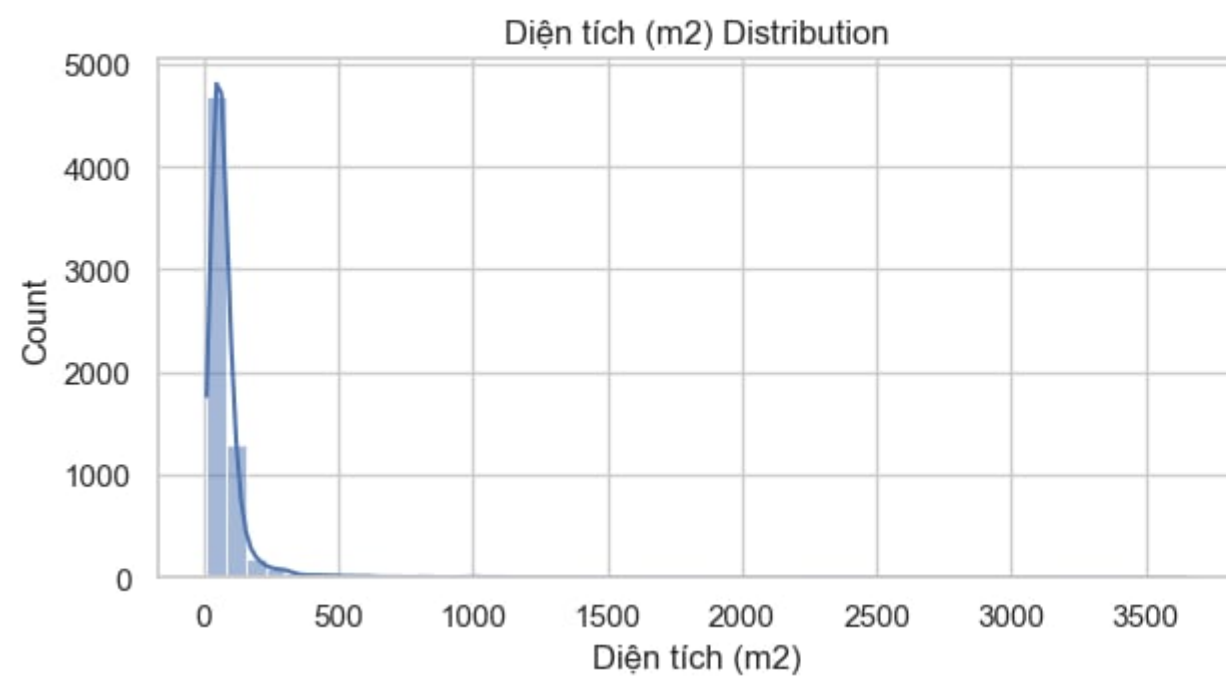
Giá trị bị thiếu

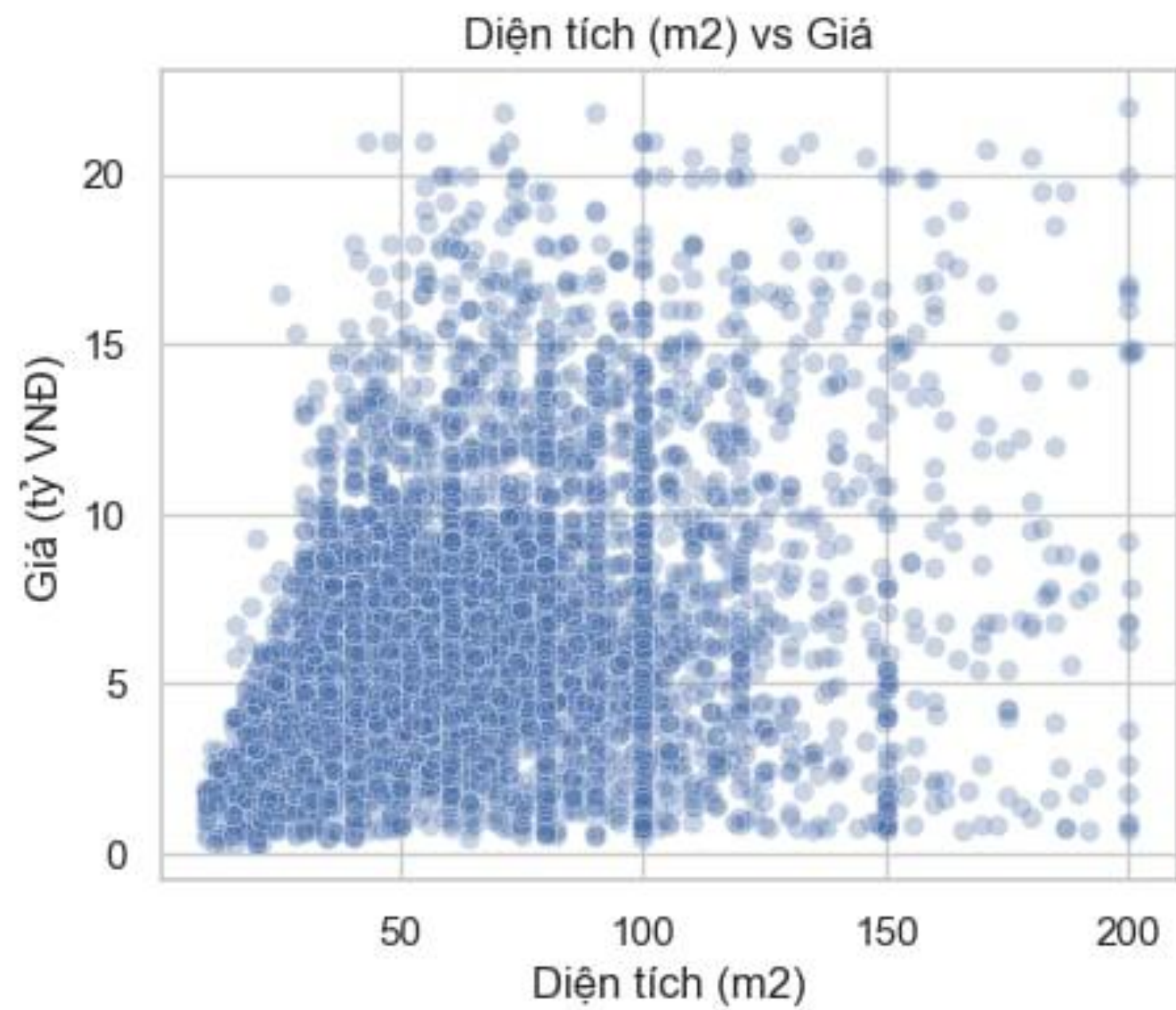
Các đặc trưng như Hướng và Nội thất có giá trị bị thiếu, được coi là thông tin quan trọng.

Biểu đồ minh họa phân bố giá nhà (phân bố dài) và sự hiện diện của các giá trị ngoại lai.



Diện tích (m2) | Outliers: 419 (6.55%)





Chiến lược Tiền xử lý: "Làm sạch trước, Chuyển đổi sau"

Cách tiếp cận hai bước của chúng tôi đảm bảo dữ liệu được chuẩn bị tốt để mô hình hóa.

01

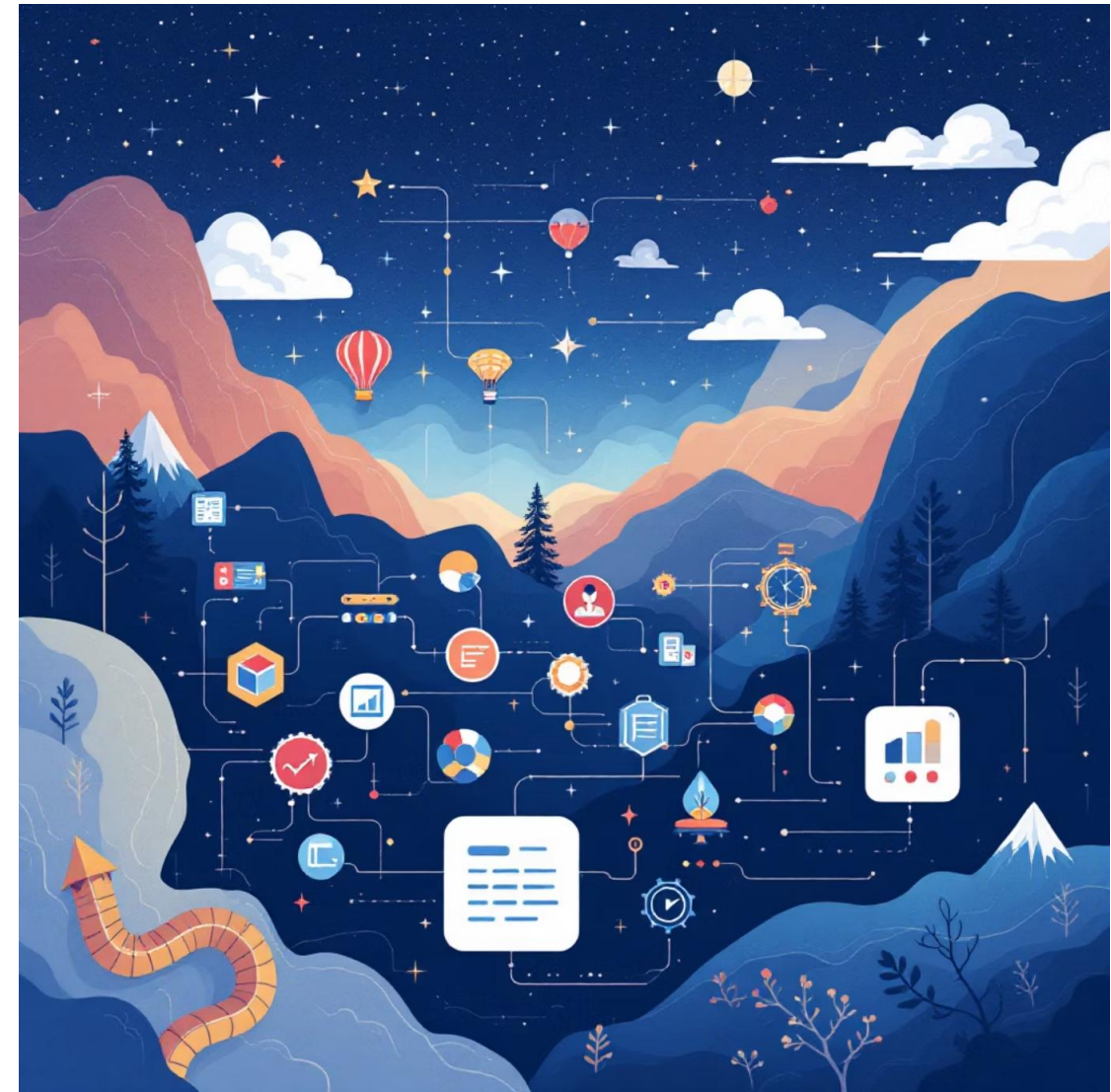
Giai đoạn 1: Làm sạch dữ liệu

Xử lý các giá trị NaN bằng cách coi chúng như thông tin (chiến lược Min/Max), thay vì loại bỏ chúng.

02

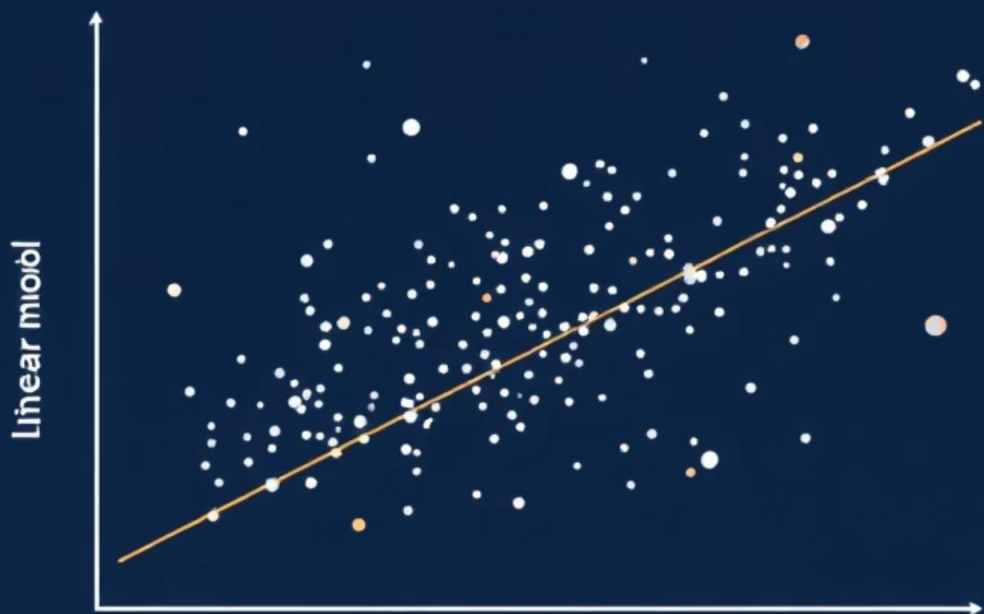
Giai đoạn 2: Kỹ thuật đặc trưng

Tạo các đặc trưng mới như "Tổng số phòng",... từ dữ liệu thô để tăng cường khả năng dự đoán.



Quy trình tiền xử lý dữ liệu từ làm sạch đến kỹ thuật đặc trưng.





linear models



Tree-based models

Lựa chọn Mô hình: Ưu tiên cây quyết định

Các mô hình dựa trên cây đã được chọn vì khả năng xử lý tốt các đặc tính phức tạp của dữ liệu bất đồng sản.



Phi tuyến tính

Mô hình cây xử lý hiệu quả các mối quan hệ phi tuyến tính phức tạp trong dữ liệu giá nhà.



Mạnh mẽ với giá trị ngoại lai

Ít nhạy cảm với các giá trị ngoại lai so với các mô hình tuyến tính, phù hợp với phân bố dài của dữ liệu.



Đa dạng

Các mô hình như Random Forest, LightGBM và CatBoost mang lại hiệu suất mạnh mẽ và khả năng khái quát hóa.

Tối ưu hóa Mô hình: Optuna

Để đạt được hiệu suất tốt nhất, chúng tôi đã sử dụng Optuna để điều chỉnh siêu tham số một cách hiệu quả và tự động.

- **Optuna:** Khung tối ưu hóa siêu tham số tự động.
- **Ưu điểm:** Hiệu quả hơn GridSearch, khám phá không gian siêu tham số tốt hơn.
- **Mục tiêu:** Tối thiểu hóa lỗi dự đoán và cải thiện độ chính xác của mô hình.

Optuna sử dụng các chiến lược lấy mẫu thông minh để tìm ra tập hợp siêu tham số tối ưu, rút ngắn đáng kể thời gian phát triển.



Tìm hiểu sâu: CatBoost

CatBoost nổi bật là mô hình hiệu quả nhất do các tính năng độc đáo của nó.

Xử lý đặc trưng phân loại

CatBoost xử lý trực tiếp các đặc trưng phân loại mà không cần mã hóa trước, giảm thiểu mất thông tin.



Xử lý giá trị bị thiếu

Khả năng tự động xử lý giá trị bị thiếu một cách hiệu quả, phù hợp với chiến lược tiền xử lý của chúng tôi.

Hiệu suất cao

Đạt được độ chính xác hàng đầu trong khi duy trì hiệu quả tính toán.



Kết quả Đánh giá & Kết luận

So sánh Hiệu suất Mô hình

Random Forest	0.6543
LightGBM	0.6789
CatBoost	0.6911

CatBoost đã thể hiện hiệu suất vượt trội, đạt R^2 là 0.6911, chứng tỏ độ chính xác cao trong việc dự đoán giá nhà.

Thành công của Quy trình

- ### Quy trình hiệu quả

Đã xây dựng thành công một quy trình dự đoán giá nhà từ đầu đến cuối.
- ### Dự đoán chính xác

Cung cấp các dự đoán giá đáng tin cậy cho thị trường bất động sản.
- ### Hệ thống mạnh mẽ

Một nền tảng vững chắc cho việc định giá bất động sản dựa trên dữ liệu.