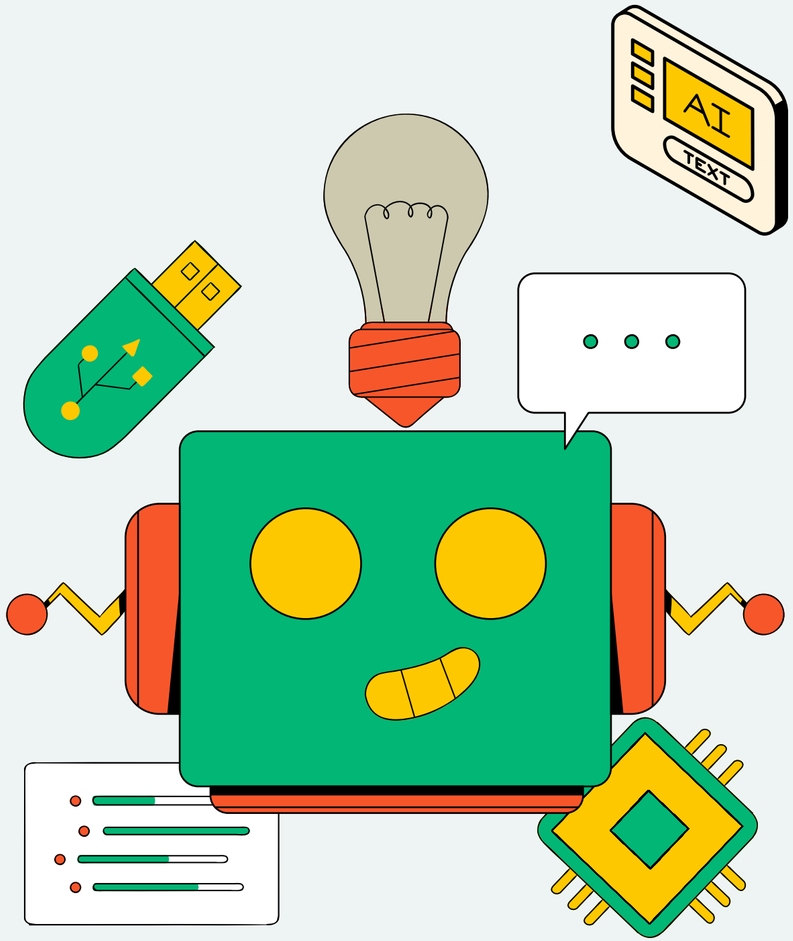
MACHINE LEARNING



DỰ ĐOÁN GIÁ NHÀ CALIFORNIA

BẰNG LINEAR REGRESSION – SVR – RANDOM FOREST

SINH VIÊN TRÌNH BÀY

NGUYỄN MINH THÁI
ĐỖ ĐỨC THỊNH

1. LINEAR REGRESSION

Linear Regression là một trong những thuật toán học máy đơn giản và cơ bản nhất thuộc nhóm hồi quy (regression). Mục tiêu của nó là tìm ra một mối quan hệ tuyến tính giữa một biến phụ thuộc (biến mục tiêu - y) và một hoặc nhiều biến độc lập (đặc trưng - X).

Mục tiêu:

-Dự đoán giá trị trung bình của nhà dựa trên các đặc trưng đầu vào như longitude, latitude,median_income, total_rooms, housing_median_age, v.v.

-Tìm ra một mối quan hệ tuyến tính giữa các đặc trưng này và median_house_value.

-Cung cấp một đường cơ sở (baseline) hiệu suất để so sánh với các mô hình phức tạp hơn như SVR và Random Forest.

2. SUPPORT VECTOR REGRESSION

SVR là một phần mở rộng của thuật toán Support Vector Machine (SVM), được thiết kế để giải quyết các bài toán hồi quy (thay vì phân loại).

Mục tiêu:

-Dự đoán median_house_value tương tự như Linear Regression.

-Xử lý các mối quan hệ phức tạp, phi tuyến tính: Với dữ liệu giá nhà, mối quan hệ giữa các đặc trưng (thu nhập, vị trí, tuổi nhà) và giá nhà không hẳn là hoàn toàn tuyến tính. SVR, đặc biệt khi sử dụng kernel phi tuyến tính (rbf trong trường hợp này), có tiềm năng nắm bắt được những mối quan hệ phức tạp hơn mà Linear Regression có thể bỏ qua.

3. RANDOM FOREST

Random Forest là một thuật toán học máy mạnh mẽ và linh hoạt thuộc nhóm học máy ensemble (ensemble learning), cụ thể là phương pháp bagging (Bootstrap Aggregating).

Mục tiêu:

-Dự đoán median_house_value với độ chính xác cao nhất có thể.

-Nắm bắt các mối quan hệ phi tuyến tính và phức tạp: Dữ liệu giá nhà thường có các mối quan hệ phức tạp, không chỉ tuyến tính. Random Forest, với bản chất là kết hợp của các cây quyết định, có khả năng học được các tương tác phức tạp và phi tuyến tính giữa các đặc trưng.

GIỚI THIỆU BÀI TOÁN

Dữ liệu chứa:

- Thu nhập vùng
- Tuổi khu nhà
- Số phòng, phòng ngủ
- Mật độ dân số
- Gần biển / xa biển

Mục tiêu:

→ Dự đoán giá nhà (median_house_value)

MÔ TẢ DATASET

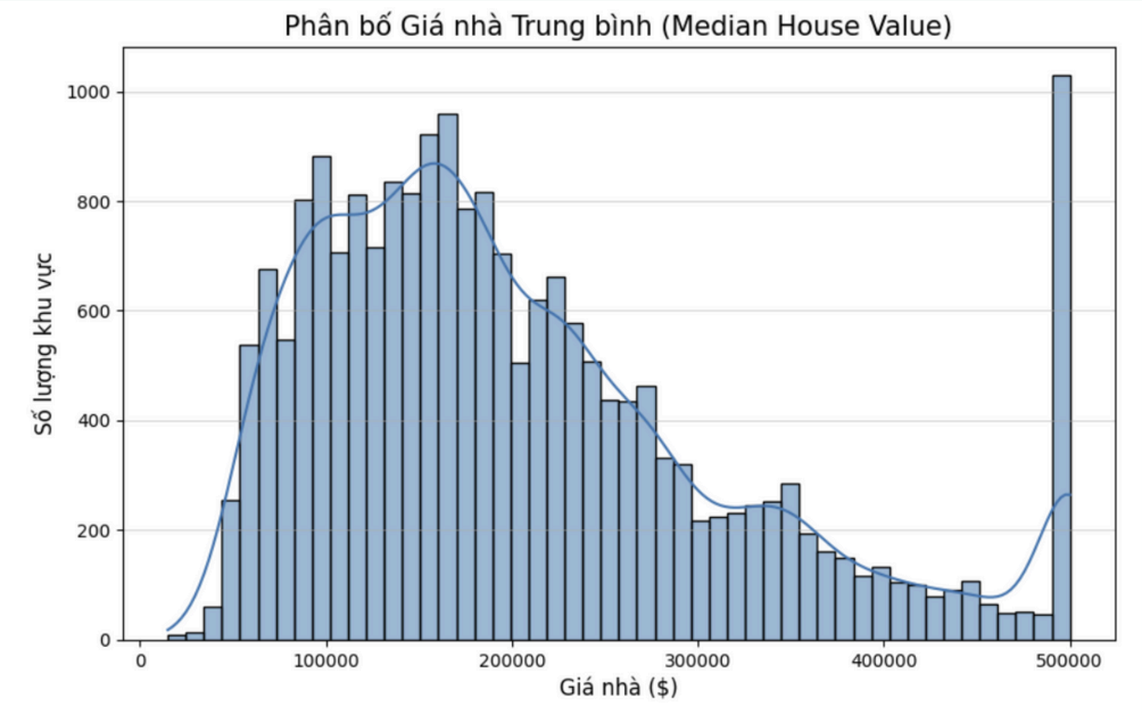
Bộ dữ liệu gồm nhiều cột mô tả thông tin dân cư và đặc điểm nhà ở. Dữ liệu có hơn 20.000 dòng, mỗi dòng đại diện cho một cụm hộ dân.

Các đặc trưng quan trọng bao gồm:

- median_income: Thu nhập trung vị (ảnh hưởng mạnh đến giá nhà)
- housing_median_age: Tuổi trung vị của các căn nhà
- total_rooms, total_bedrooms: Tổng số phòng và phòng ngủ
- population, households: Dân cư và số hộ gia đình
- ocean_proximity: Vị trí gần biển (yếu tố quan trọng trong bất động sản)

Biến mục tiêu cần dự đoán là median_house_value – giá nhà trung vị của mỗi khu vực.

KIỂM TRA PHÂN BỐ (DISTRIBUTION)



Nhận xét:

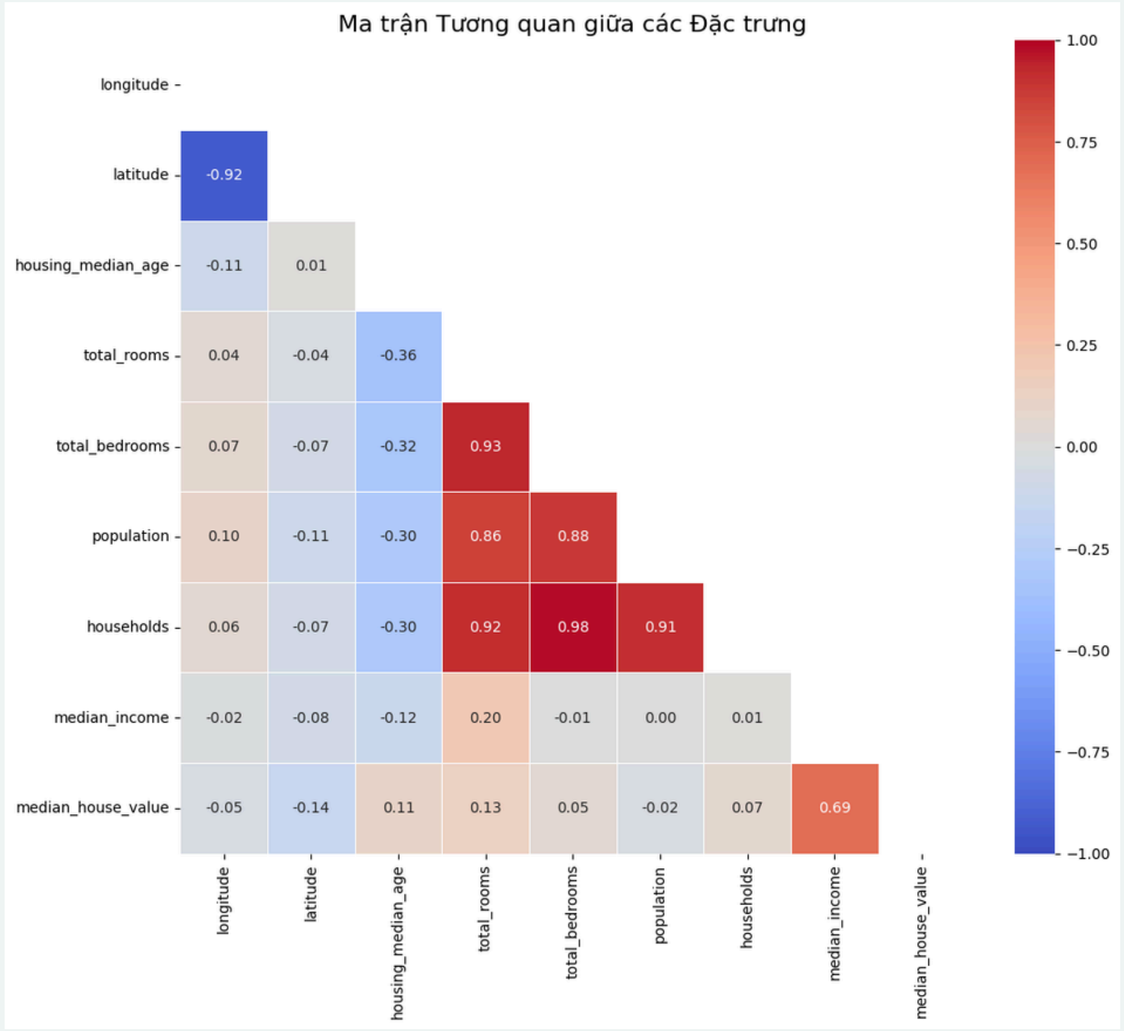
- Không phân bố đều
- Có trần giá 500.000 → cần lưu ý trong ML
- Bất đối xứng nhẹ

MA TRẬN TƯƠNG QUAN (HEATMAP)

Yếu tố ảnh hưởng mạnh:

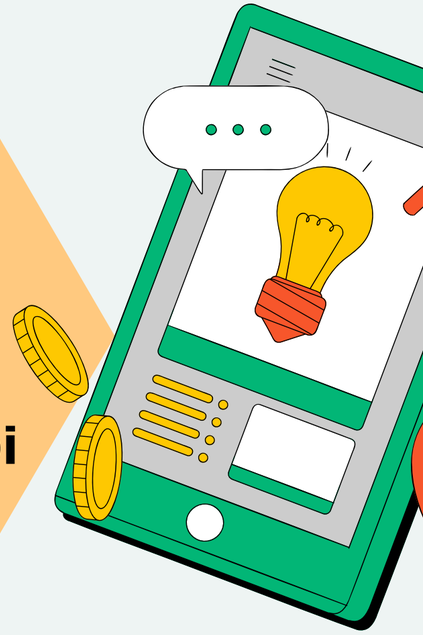
Median Income

```
Mức độ tương quan với Giá nhà (Median House Value):
median_house_value    1.000000
median_income         0.688075
total_rooms           0.134153
housing_median_age    0.105623
households            0.065843
total_bedrooms        0.049686
population           -0.024650
longitude            -0.045967
latitude             -0.144160
Name: median_house_value, dtype: float64
```



TIỀN XỬ LÝ (PREPROCESSING)

Chuẩn hóa dữ liệu (StandardScaler):
Áp dụng cho Linear Regression và đặc biệt quan trọng với SVR, giúp mô hình hội tụ tốt hơn.

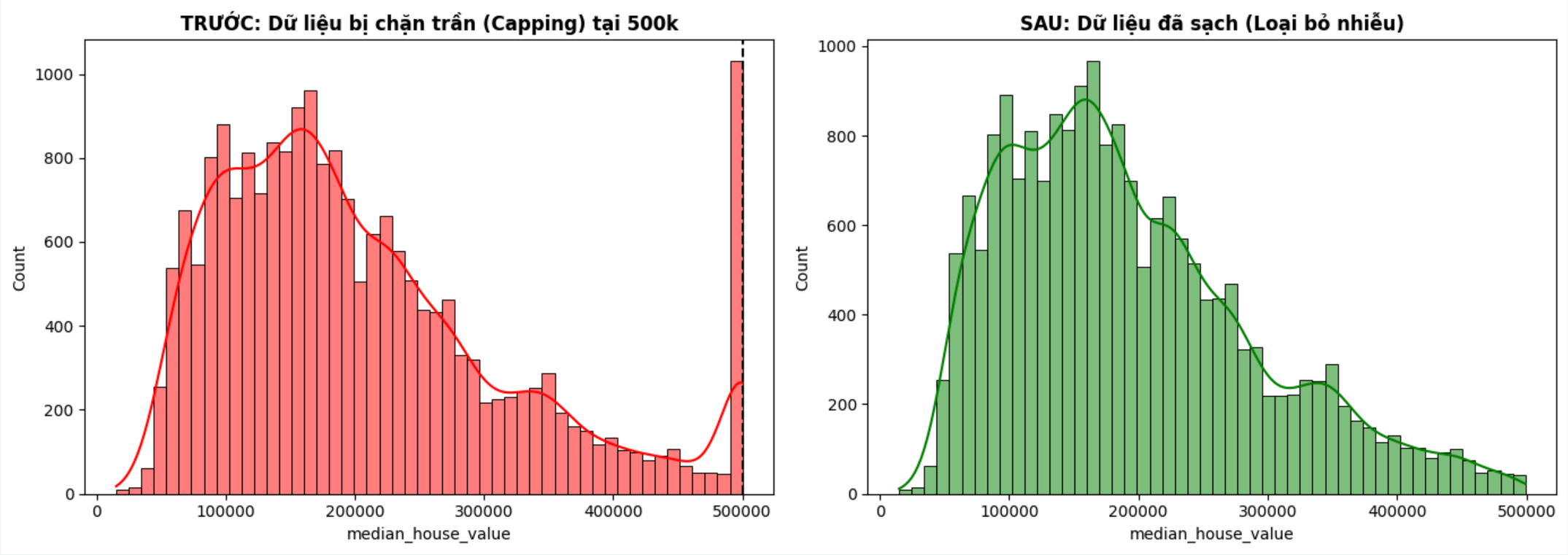


One-hot Encoding:
Biến phân loại ocean_proximity được chuyển sang dạng số để phù hợp với mô hình.

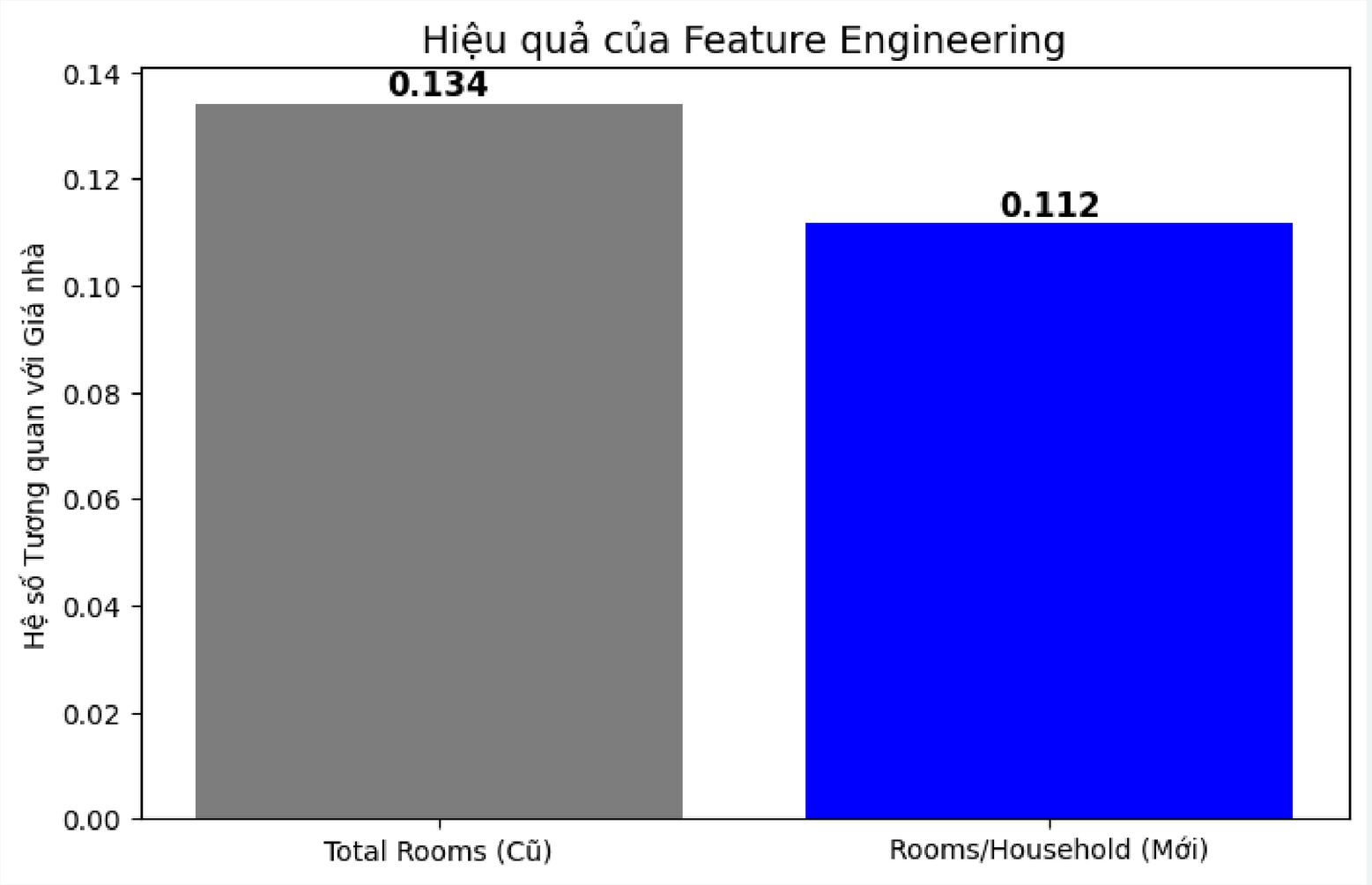
Xử lý dữ liệu thiếu (Missing Values):
Các giá trị thiếu trong cột total_bedrooms được thay thế bằng median của cột để tránh làm mất dữ liệu.

Chia dữ liệu train/test (80/20):
Đảm bảo đánh giá mô hình khách quan và không bị overfitting.

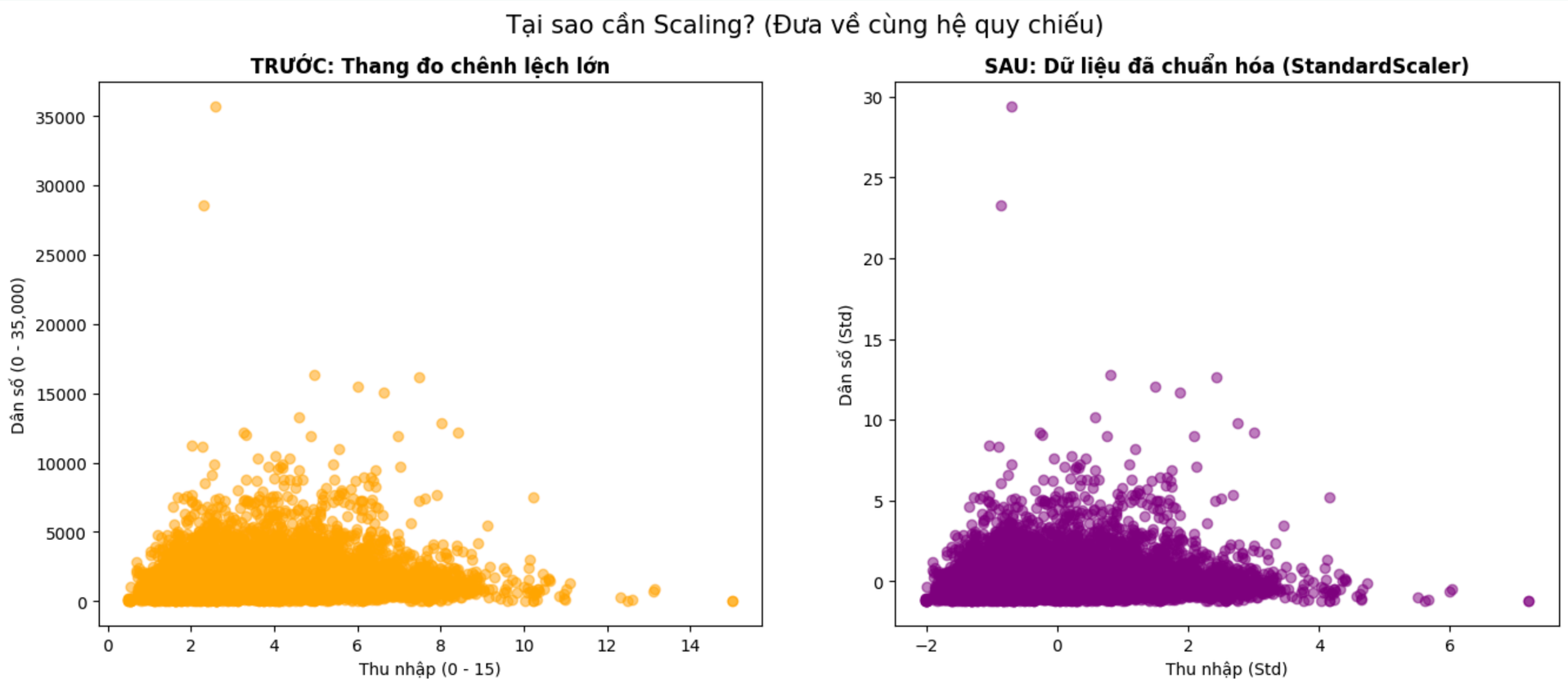
LOẠI BỎ DỮ LIỆU NHIỄU (CAPPING REMOVAL)



KỸ THUẬT ĐẶC TRƯNG (FEATURE ENGINEERING)



CHUẨN HÓA DỮ LIỆU (FEATURE SCALING) – QUAN TRỌNG CHO SVR



CÁC MÔ HÌNH ĐƯỢC SỬ DỤNG

```
--- Đánh giá Linear Regression ---
RMSE (Sai số trung bình): $60,512.44
R2 Score (Độ chính xác): 0.6181
------------------------------
--- Đánh giá Support Vector Regression (SVR) ---
RMSE (Sai số trung bình): $99,139.79
R2 Score (Độ chính xác): -0.0252
------------------------------
--- Đánh giá Random Forest Regressor ---
RMSE (Sai số trung bình): $46,522.14
R2 Score (Độ chính xác): 0.7743
------------------------------
```


```
--- Đánh giá Linear Regression ---
RMSE (Sai số trung bình): $60,512.44
R2 Score (Độ chính xác): 0.6181
------------------------------
--- Đánh giá Support Vector Regression (SVR) ---
RMSE (Sai số trung bình): $99,139.79
R2 Score (Độ chính xác): -0.0252
------------------------------
--- Đánh giá Random Forest Regressor ---
RMSE (Sai số trung bình): $46,522.14
R2 Score (Độ chính xác): 0.7743
------------------------------
```


1. Linear Regression
   - Là mô hình hồi quy tuyến tính cơ bản, dễ triển khai và diễn giải.
   - Phù hợp khi mối quan hệ giữa input và output gần tuyến tính.
   - Tốc độ nhanh nhưng độ chính xác hạn chế.
2. Support Vector Regression (SVR – RBF Kernel)
   - Có khả năng mô hình hóa quan hệ phi tuyến nhờ kernel.
   - Nhạy mạnh với chuẩn hóa dữ liệu.
   - Cho kết quả tốt hơn Linear Regression nhưng thời gian huấn luyện lâu.
3. Random Forest Regression
   - Tập hợp nhiều cây quyết định, giúp giảm overfitting.
   - Hoạt động tốt với dữ liệu phi tuyến hoặc nhiều nhiễu.
   - Thường đạt hiệu năng cao nhất trong các bài toán tabular.

KẾT QUẢ MÔ HÌNH

	Mô hình	MAE (Sai số tuyệt đối)	MSE (Bình phương sai số)	R2 Score (Độ chính xác)
0	Linear Regression	$44,825.33	3,661,755,729.56	0.6181
1	SVR	$76,485.45	9,828,697,621.07	-0.0252
2	Random Forest	$31,027.62	2,164,309,717.87	0.7743

$$R^2 = 1 - \frac{\Sigma_i^N (y_i - \hat{y}_i)^2}{\Sigma_i^N (y_i - \bar{y})^2}$$

$$= 1 - \frac{\text{Sai số của mô hình}}{\text{Độ lệch so với trung bình dữ liệu}}$$

=>"Mô hình đã mô phỏng được 77% mức biến thiên của thị trường giá nhà thực tế."

R2 là phép so sánh xem mô hình của em giảm được bao nhiêu lỗi so với cách đoán mò bằng trung bình

ĐÁNH GIÁ & NHẬN XÉT

Từ kết quả đánh giá:

Linear Regression: Dễ dùng, nhanh, nhưng không mô tả tốt dữ liệu phi tuyến.

SVR: Hiệu quả hơn Linear Regression nhờ kernel phi tuyến, nhưng tốc độ chậm và phụ thuộc mạnh vào scaling.

Random Forest: Cho kết quả chính xác nhất, ổn định và không cần scaling. Đây là mô hình được đề xuất cho bài toán.

KẾT LUẬN

Random Forest cho hiệu năng tốt nhất khi dự đoán giá nhà California.

Yếu tố quan trọng nhất ảnh hưởng đến giá nhà là median_income.

Phân bố dữ liệu lệch và trần giá 500.000 đã ảnh hưởng đến mô hình, cần được lưu ý khi mở rộng bài toán.

Bộ dữ liệu sau khi xử lý giúp mô hình đạt kết quả ổn định và đáng tin cậy.

HƯỚNG PHÁT TRIỂN

01

Thử nghiệm thêm các mô hình mạnh hơn như XGBoost, LightGBM.

02

Tối ưu tham số (GridSearchCV/RandomizedSearch).

03

Loại bỏ outlier để cải thiện Linear Regression.

04

Sử dụng tọa độ địa lý để vẽ bản đồ giá nhà trực quan.

05

Thử thêm kỹ thuật Feature Engineering nhằm tăng độ chính xác.


Q &
A