Dự đoán giá nhà bằng Linear Regression và Lasso Regression

Giảng viên hướng dẫn: PGS.TS. Nguyễn Văn Hậu

Người trình bày: Lê Quang Trung

Nội dung

- Tổng quan bài toán
- Mô hình học máy
- Cài đặt mô hình
- Demo chương trình

Bài toán dự đoán giá nhà

- Là 1 trong những bài toán liên quan đến việc quyết định mua bán đối với khách hàng

- Đối với các công ty bất động sản: Phân tích thị trường, tìm kiếm cơ hội đầu tư, và hỗ trợ thẩm định giá tài sản.

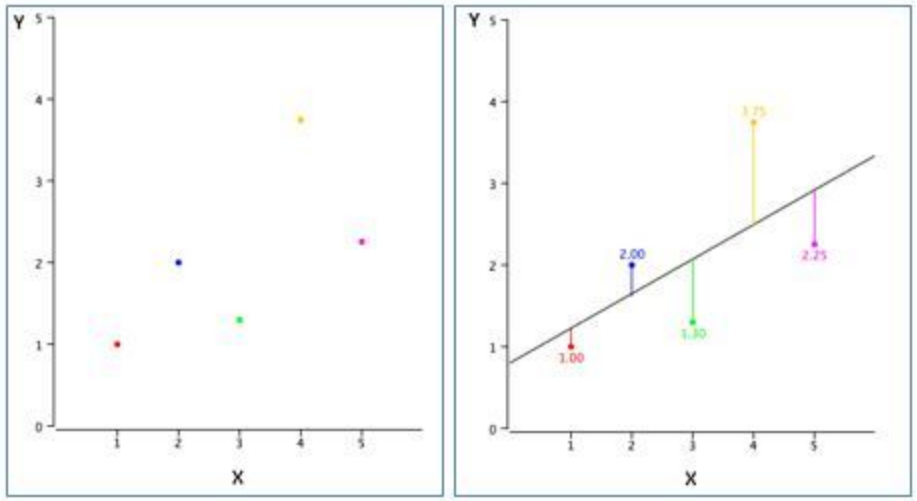
=> Bài toán dự đoán giá nhà thuộc loại học có giám sát (supervised learning) và cụ thể là bài toán hồi quy (regression)

Các mô hình được sử dụng

- Linear Regression
- Lasso Regression

Linear Regression

- Định nghĩa
  - Hồi quy tuyến tính (Linear Regression) là một mô hình thống kê/machine learning dùng để dự đoán giá trị của biến phụ thuộc (y) dựa trên một hoặc nhiều biến độc lập (x) bằng cách tìm một đường thẳng (hoặc siêu phẳng) sao cho mô tả tốt nhất mối quan hệ giữa chúng.

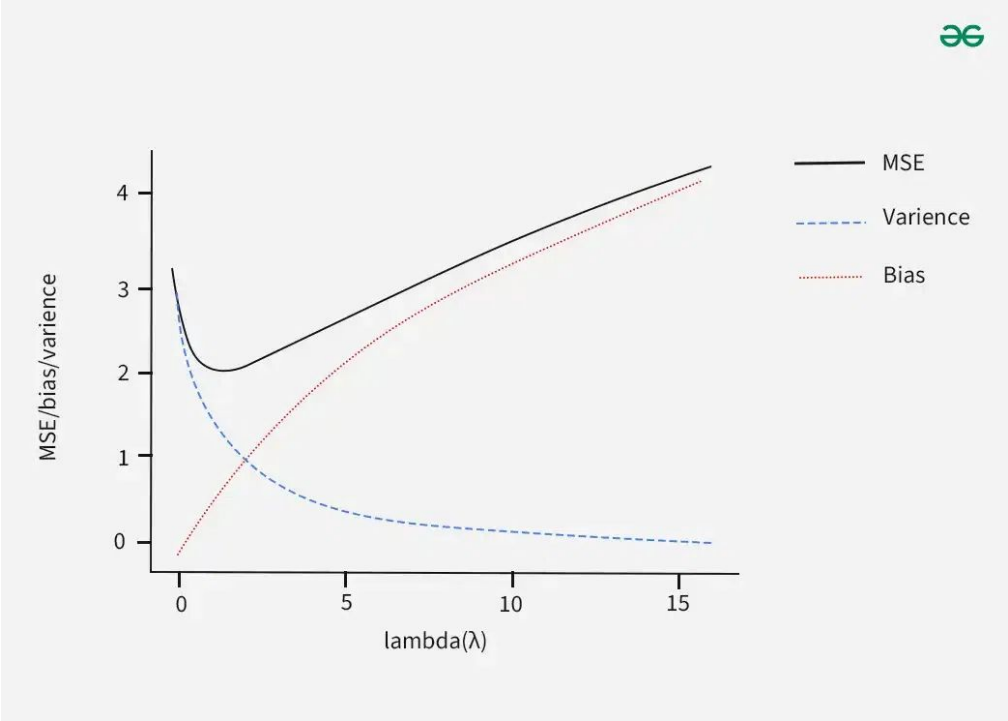


Linear Regression

- Công thức: wx + b = y
  - **W: Trọng số**
  - b: độ lệch
  - x: đầu vào của mô hình
  - y: đầu ra của mô hình

Lasso Regression

- Hồi quy Lasso (Lasso Regression) là một dạng hồi quy tuyến tính có regularization, sử dụng ràng buộc L1 để vừa dự đoán tốt vừa tự động chọn lọc đặc trưng (feature selection).



Lasso Regression

- Lasso tìm ra mô hình tuyến tính tối ưu bằng cách ép nhiều trọng số về 0 để chọn những feature quan trọng nhất.

$$L_{lasso} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - X_i w\right)^2 + \lambda \sum |w_j|$$

MSE

Forces some weights to be exactly zero

Các Metrics đánh giá

- MSE: Mean Squared Error
- RMSE: Root Mean Squared Error
- R2 : R-squared

Mô tả dữ liệu

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00
5	Super built-up Area	Ready To Move	Whitefield	2 BHK	DuenaTa	1170	2.0	1.0	38.00
6	Super built-up Area	18-May	Old Airport Road	4 BHK	Jaades	2732	4.0	NaN	204.00
7	Super built-up Area	Ready To Move	Rajaji Nagar	4 BHK	Brway G	3300	4.0	NaN	600.00
8	Super built-up Area	Ready To Move	Marathahalli	3 BHK	NaN	1310	3.0	1.0	63.25
9	Plot Area	Ready To Move	Gandhi Bazar	6 Bedroom	NaN	1020	6.0	NaN	370.00
10	Super built-up Area	18-Feb	Whitefield	3 BHK	NaN	1800	2.0	2.0	70.00
11	Plot Area	Ready To Move	Whitefield	4 Bedroom	Prrry M	2785	5.0	3.0	295.00
12	Super built-up Area	Ready To Move	7th Phase JP Nagar	2 BHK	Shncyes	1000	2.0	1.0	38.00
13	Built-up Area	Ready To Move	Gottigere	2 BHK	NaN	1100	2.0	2.0	40.00

https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data

Mô tả dữ liệu

Tổng quan các trường trong dataset

area_type : Loại bất động sản

availability : Tình trạng của bất động sản

location : Vị trí, khu vực của bất động sản

size : Quy mô, kích thước thường là số phòng ngủ cho 1 căn nhà

society : Tên tổ dân phố nơi toạ lạc của các căn nhà

total_sqft : Tổng diện tích tính bằng đơn vị feet vuông

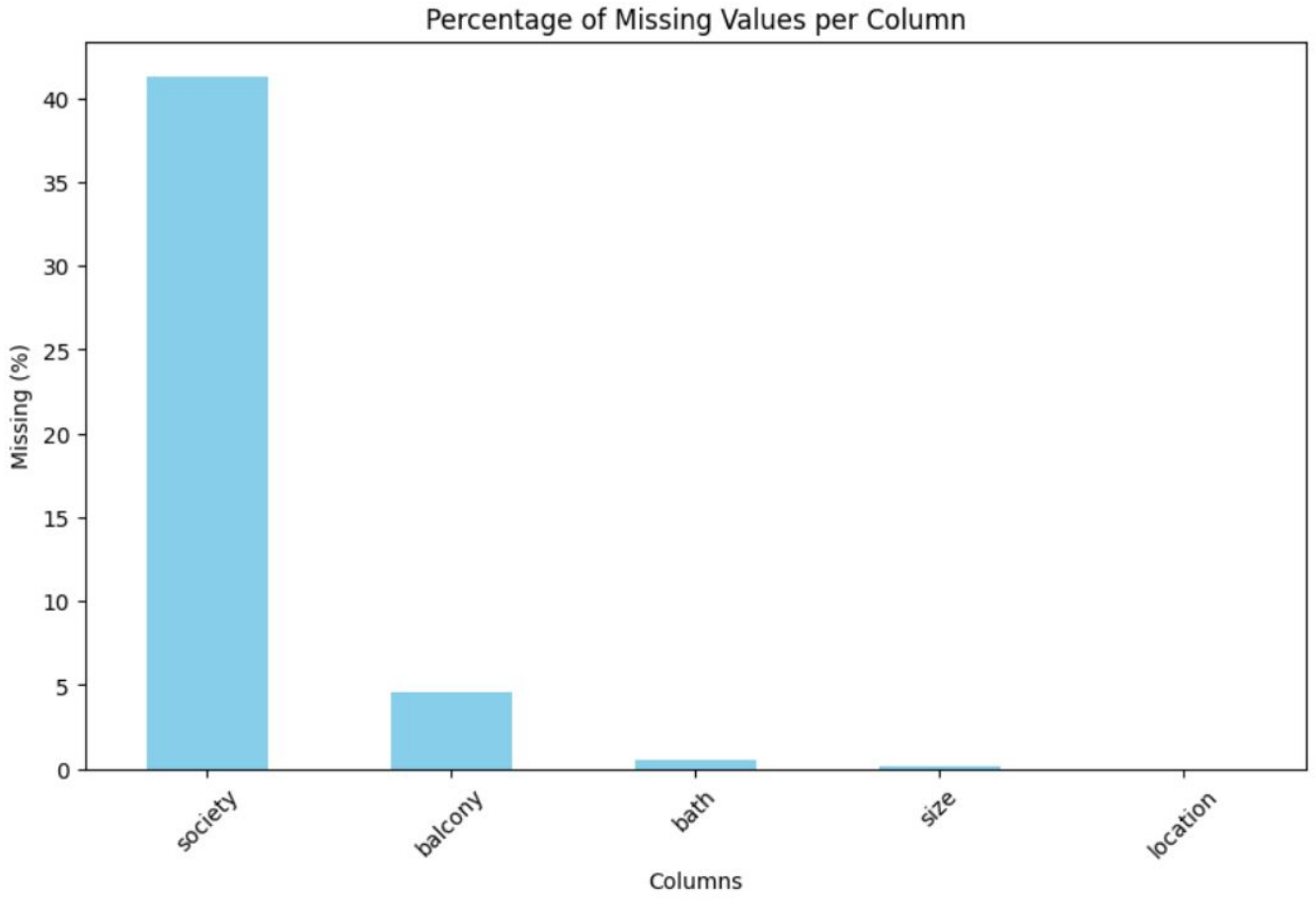
bath : Số lượng phòng tắm của các căn nhà

balcony : Số lượng ban công

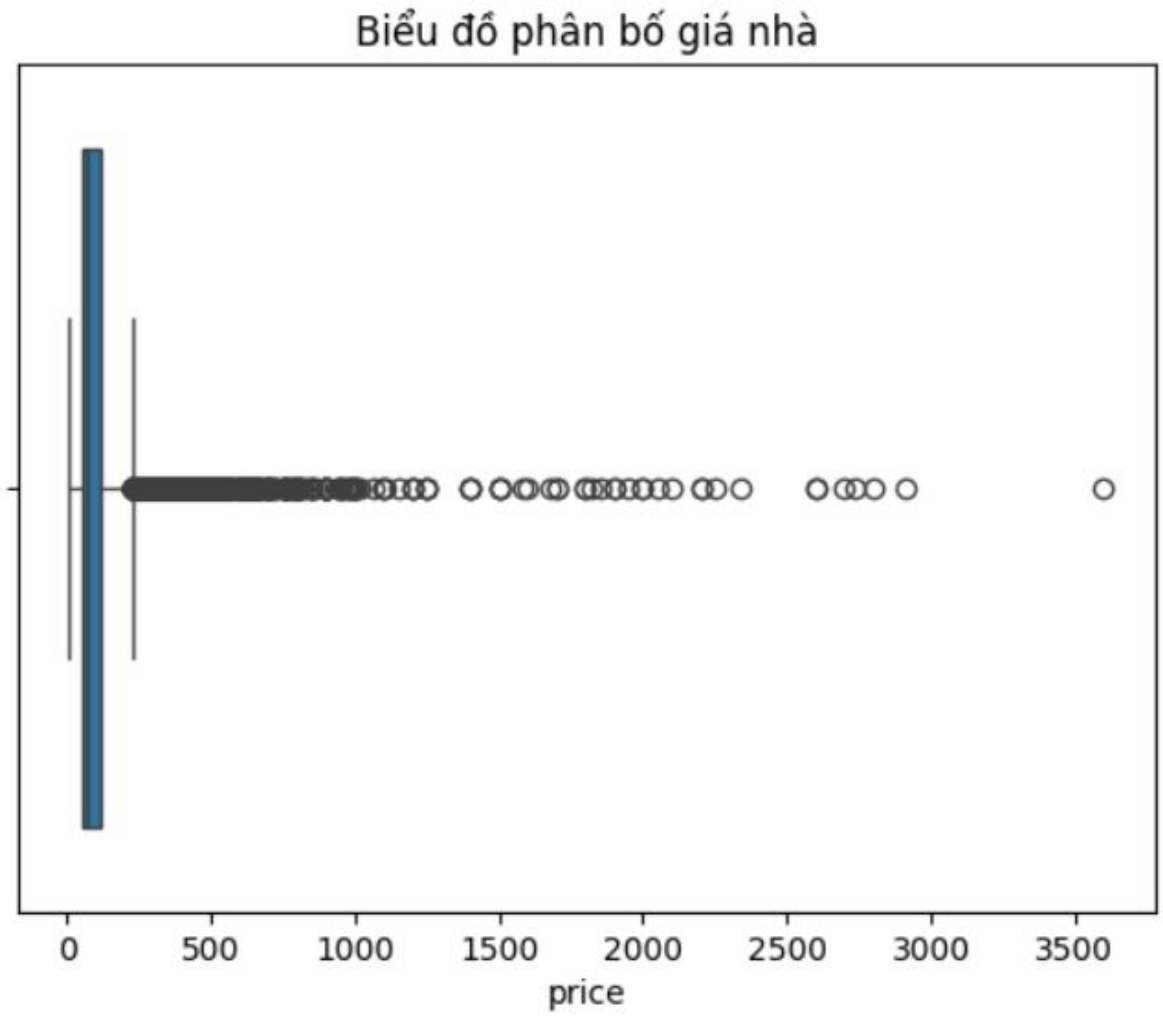
price : Glá bán tính bằng đơn vị Lakh

Phân tích dữ liệu

- Thống kê số lượng dữ liệu bị khuyết thiếu



Phân tích dữ liệu



Tiền xử lí dữ liệu

- Xoá những feature không cần thiết

```
# Xoá các cột không cần thiết
df.drop(columns= ['area_type', 'availability', 'society', 'balcony'], inplace = True)
```

- Điền đầy những giá trị bị thiếu ở các cột có tỉ lệ khuyết thiếu thấp

```
df['size']=df['size'].fillna('2 BHK')
df['bath'] = df['bath'].fillna(df['bath']).median()
```

Tiền xử lí dữ liệu

- Convert sang kiểu float và xử lí chuỗi dữ liệu đặc biệt cho total_sqft

```
df['total_sqft'] = df['total_sqft'].apply(convert_sqft_to_num) # Áp dụng cho toàn bộ cột
```

- Tách chuỗi để trong cột size để lấy số lượng phòng

```
df['bhk'] = df['size'].apply(lambda x : int(x.split(' ')[0]))
df.head(20)
```

Tiền xử lí dữ liệu

- One-Hot Encoding
  - One-Hot Encoding (Mã hóa One-Hot) là một kỹ thuật tiền xử lý dữ liệu được sử dụng để **chuyển đổi dữ liệu phân loại (categorical data)** thành định dạng **số** (numerical format) mà các thuật toán học máy có thể hiểu và sử dụng được.

- Standardization
  - Tiêu chuẩn hóa là quá trình biến đổi dữ liệu sao cho các giá trị phân phối của một đặc trưng có **giá trị trung bình bằng 0** và **độ lệch chuẩn bằng 1**.

Đánh giá mô hình

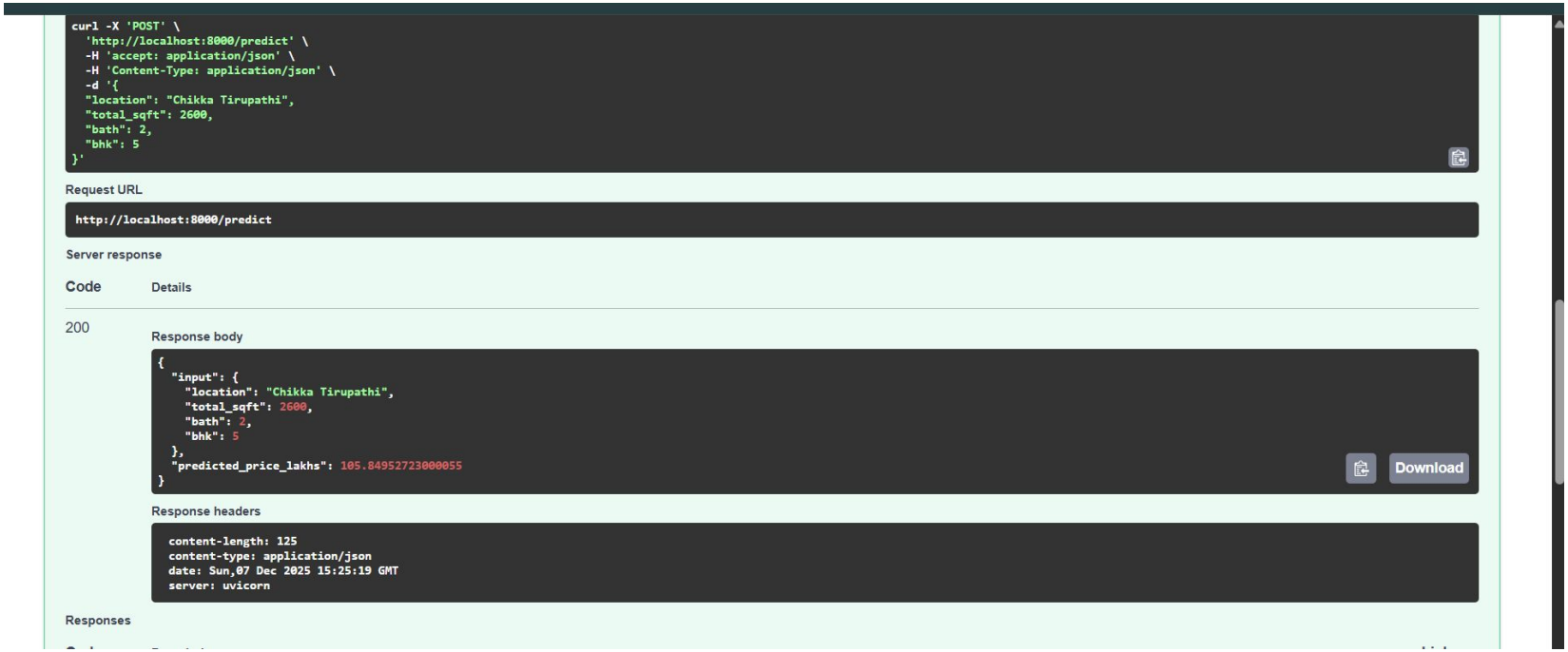
- Linear Regression

```
R2 Score 0.8627447700198984
MSE score 974.0141202397293
RMSE score 31.209199288666944
```

Đánh giá mô hình

```
R2 Score 0.8553577729692834
MSE score 1026.434996548131
RMSE score 32.03802422978251
```

Demo



```
curl -X 'POST' \
  'http://localhost:8000/predict' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "location": "Chikka Tirupathi",
  "total_sqft": 2600,
  "bath": 2,
  "bhk": 5
}'
```

Request URL

```
http://localhost:8000/predict
```

Server response

Code	Details
200	Response body

```
{
  "input": {
    "location": "Chikka Tirupathi",
    "total_sqft": 2600,
    "bath": 2,
    "bhk": 5
  },
  "predicted_price_lakhs": 105.84952723000055
}
```

Download

Response headers

```
content-length: 125
content-type: application/json
date: Sun,07 Dec 2025 15:25:19 GMT
server: uvicorn
```

Responses

Trân trọng cảm ơn!

Q&A