Bài giảng môn học:

Học máy (Machine Learning)

CHƯƠNG 3:
HỌC CÓ GIÁM SÁT – Phần 4
(Supervised Learning)

Giảng viên: Đặng Văn Nam

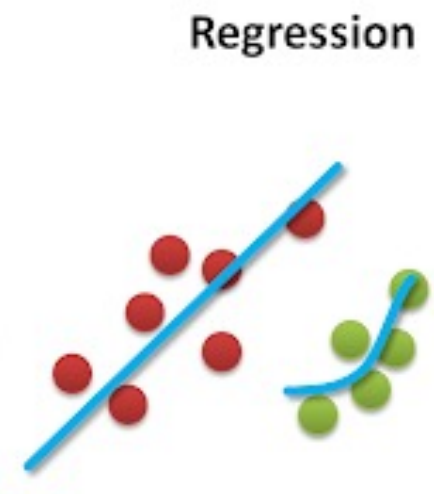
Email: dangvannam@humg.edu.vn

Nội dung chương 3 – Phần 4

Phân tích hồi quy (Regression Analysis)

1. Giới thiệu
2. Một số mô hình hồi quy cơ bản
3. Đánh giá độ chính xác của mô hình hồi quy
4. Bài tập thực hành

1. Giới thiệu bài toán hồi quy (Regression)



Bài toán hồi quy

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	**0**
4.9	3	1.4	0.2	**1**
4.7	3.2	1.3	0.2	**1**
4.6	3.1	1.5	0.2	**2**
5	3.6	1.4	0.2	**0**
5.4	3.9	1.7	0.4	**2**
4.6	3.4	1.4	0.3	**0**
5	3.4	1.5	0.2	**2**
4.4	2.9	1.4	0.2	**2**
4.9	3.1	1.5	0.1	**1**

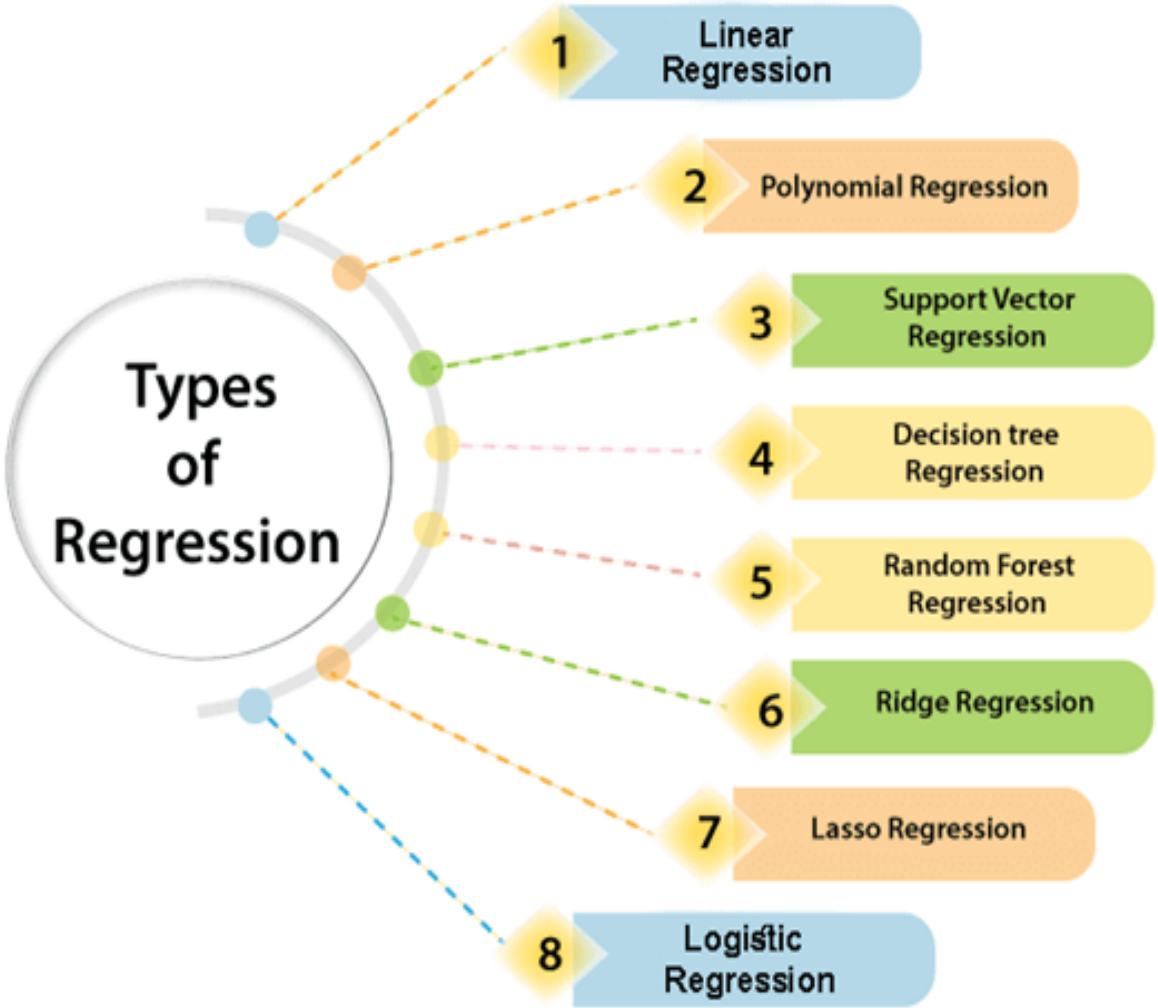
Bài toán dự đoán hóa Lan

Weight(Carat)	Price(USD)
0.23	**484**
0.31	**942**
0.20	**345**
1.02	**4459**
1.63	**14022**
1.14	**4212**
2.01	**11925**
1.28	**9548**
1.70	**11605**
1.01	**4642**

Bài toán dự đoán giá kim cương

Bài toán hồi quy

Hai nhóm bài toán cơ bản trong học có giám sát là classification (phân loại) và regression (hồi quy), trong đó biến đầu ra của bài toán phân loại có các giá trị rời rạc, hữu hạn; biến đầu ra của bài toán hồi quy có các giá trị liên tục, vô hạn.



2. Hồi quy tuyến tính (Linear Regression)

Giới thiệu

Một căn nhà rộng x_1 **m^2**, có x_2 **phòng ngủ** và cách trung tâm thành phố x_3 **km** có giá là bao nhiêu?

STT	Diện tích (m2)	Số phòng ngủ	Khoảng cách tới trung tâm thành phố (Km)	Giá bán (USD)
1	70	3	5	35,000
2	50	2	8	21,000
3	85	4	1	53,000
4	65	2	10	28,000
5	45	1	1	30,000
6	55	3	3	35,000
…	…	…	…	…

Giả sử chúng ta đã có số liệu thống kê từ 1000 căn nhà trong thành phố đó, liệu rằng khi có một căn nhà mới với các thông số về diện tích, số phòng ngủ và khoảng cách tới trung tâm, chúng ta có thể dự đoán được giá của căn nhà đó không?

Nếu có thì hàm dự đoán **y=f(x)** sẽ có dạng như thế nào?
Ở đây $x=[x_1,x_2,x_3]$ là một vector hàng chứa thông tin *input*,
y là một số vô hướng (scalar) biểu diễn *output* (tức giá của căn nhà trong ví dụ này).

Giới thiệu

STT	Diện tích (m2)	Số phòng ngủ	Khoảng cách tới trung tâm thành phố (Km)	Giá bán (USD)
1	70	3	5	35,000
2	50	2	8	21,000
3	85	4	1	53,000
4	65	2	10	28,000
5	45	1	1	30,000
6	55	3	3	35,000
…	…	…	…	…

- Một cách đơn giản nhất, chúng ta có thể thấy rằng:
  - **i) diện tích nhà càng lớn thì giá nhà càng cao;**
  - **ii) số lượng phòng ngủ càng lớn thì giá nhà càng cao;**
  - **iii) càng xa trung tâm thì giá nhà càng giảm.**

Một hàm số đơn giản nhất có thể mô tả mối quan hệ giữa giá nhà và 3 đại lượng đầu vào là:

$$y \approx f(x)$$

$$f(x) = w_1x_1 + w_2x_2 + w_3x_3 + w_0 \quad (1)$$

Trong đó: w_1, w_2, w_3, w_0 là các hằng số, w_0 còn được gọi là bias.
Mối quan hệ $y \approx f(x)$ bên trên là một mối quan hệ tuyến tính (linear).
Bài toán chúng ta đang làm là một bài toán thuộc loại hồi quy (regression). Bài toán đi tìm các hệ số tối ưu $\{w_1, w_2, w_3, w_0\}$ chính vì vậy được gọi là bài toán Linear Regression.

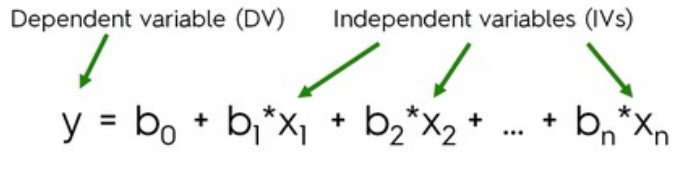
Hồi quy tuyến tính (Linear Regression).

- Hồi quy tuyến tính với 1 biến độc lập X là biến đầu vào (input) để xác định 1 biến đầu ra y (target) – **Simple Linear Regression.**
- Hồi quy tuyến tính với n biến độc lập X_1, …., X_n để xác định 1 biến đầu ra y (target) – **Multiple Linear Regression.**

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression



Linear hay *tuyến tính* hiểu một cách đơn giản là *thẳng, phẳng*.

- Không gian hai chiều, một hàm số được gọi là *tuyến tính* nếu đồ thị của nó có dạng một *đường thẳng*.
- Không gian ba chiều: một hàm số được goi là *tuyến tính* nếu đồ thị của nó có dạng một *mặt phẳng*.
- Không gian nhiều hơn 3 chiều, khái niệm *mặt phẳng* không còn phù hợp nữa, thay vào đó, một khái niệm khác ra đời được gọi là *siêu mặt phẳng* (*hyperplane*).

Simple Linear Regression.

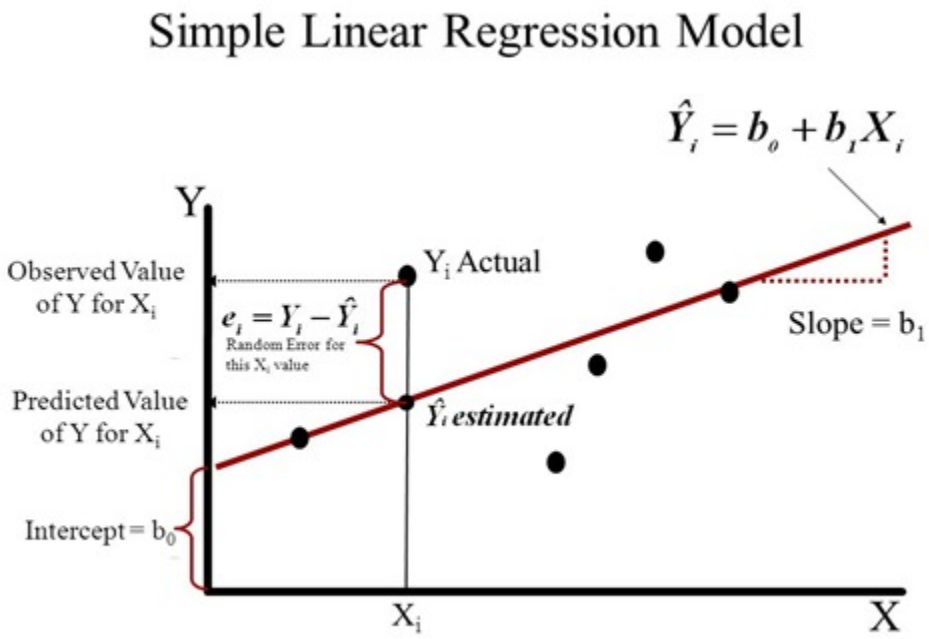
Hồi quy tuyến tính với 1 biến độc lập X là biến đầu vào (input) để xác định 1 biến đầu ra (target)

→ Xác định phương trình:

y = f(x)

$$\hat{y} = \beta_0 + \beta_1 X$$

target — coefficients — input



Mục tiêu ước lượng các tham số b_i sao cho sai số nhỏ nhất.

$$RSS = \sum_{i}^{n} (y_i - \hat{y}_i)^2$$

- **y:** Giá trị thật trong tập train (Outcome).
- **ŷ:** Giá trị mà mô hình linear regression dự đoán được.

Multiple Linear Regression.

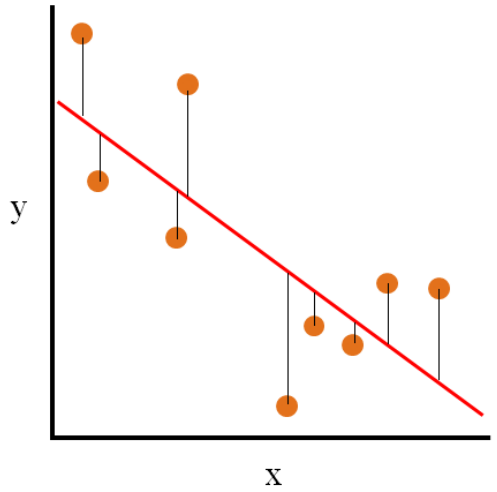
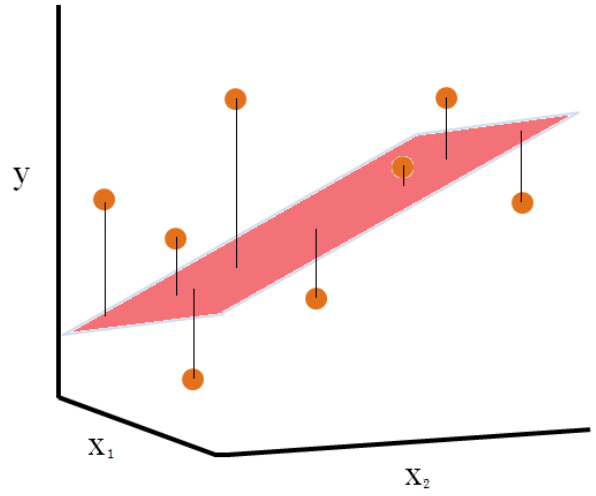
Hồi quy tuyến tính với n biến độc lập (X_1, X_2 …. X_n)

$$\hat{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

coefficients

target

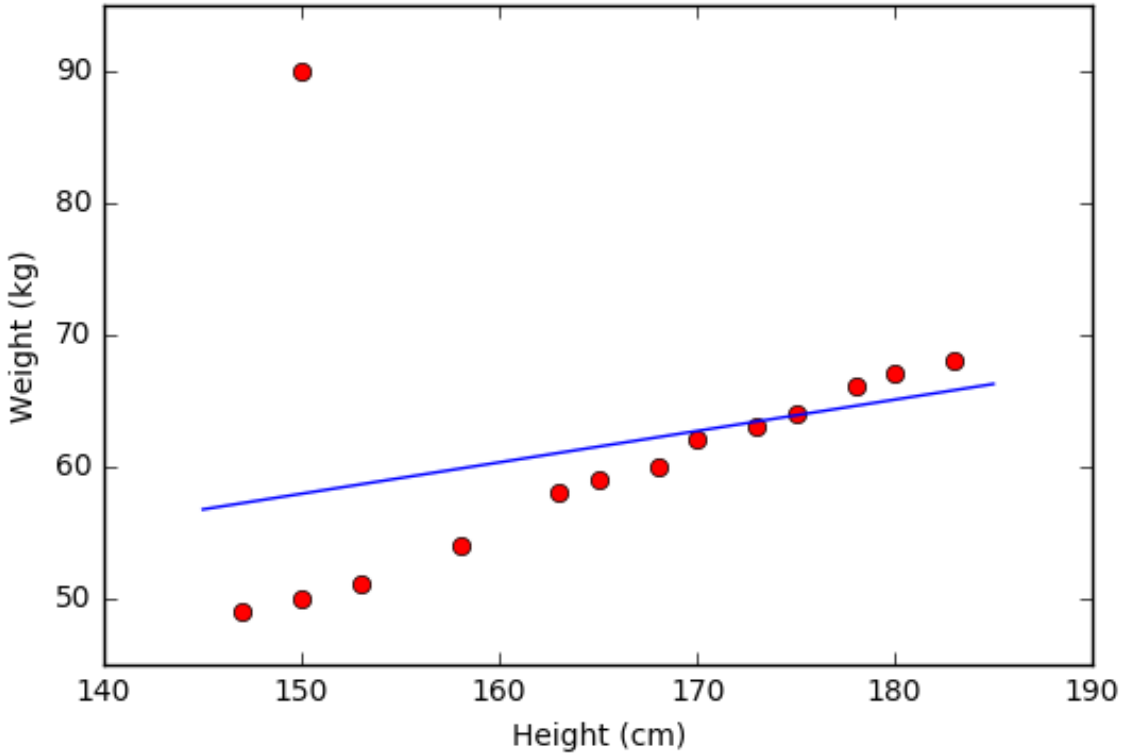
inputs

Simple Linear Regression	Multiple Linear Regression (2 Independent Variables (x_1, x_2))
	

Nhược điểm của hồi quy tuyến tính

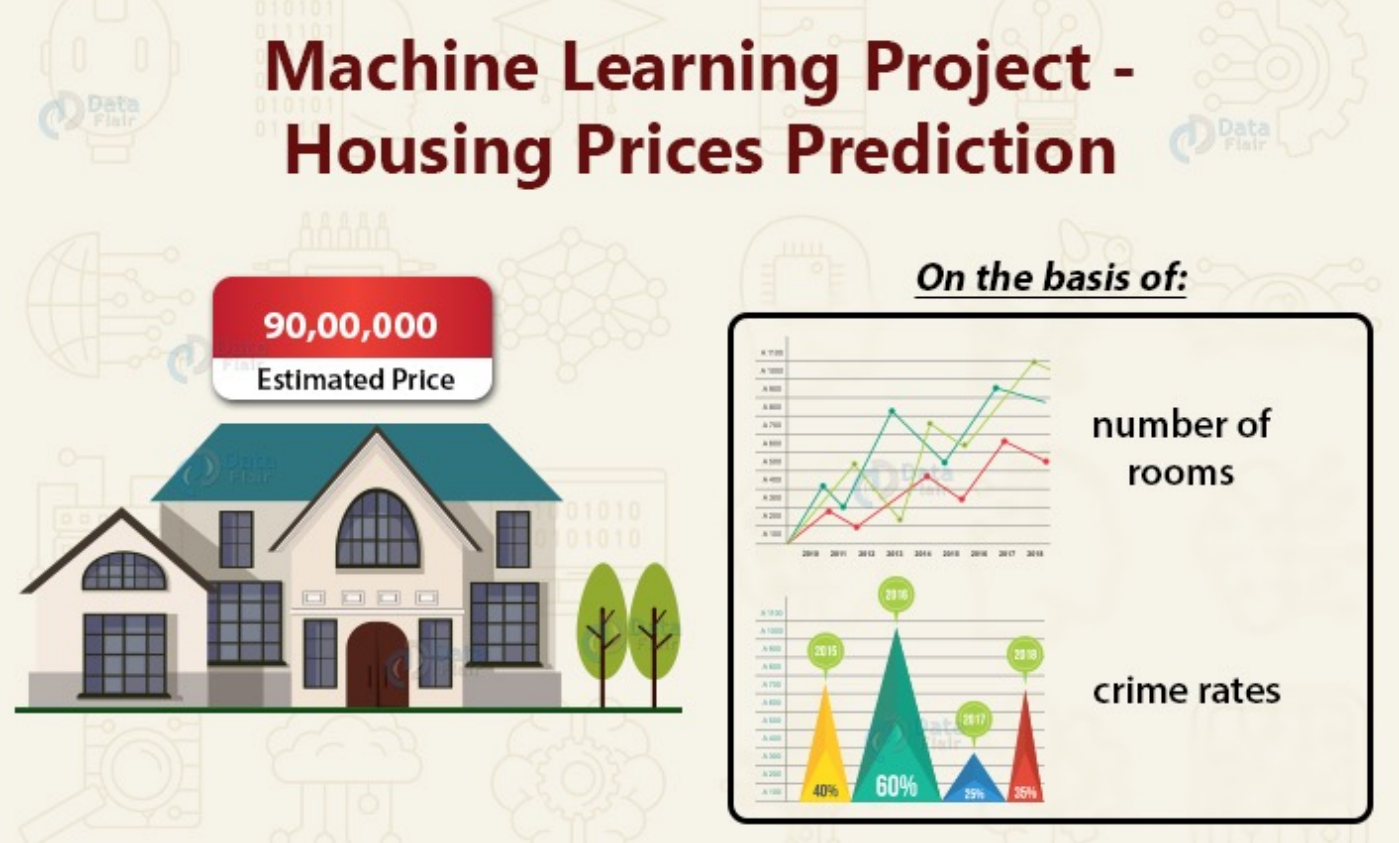
- Linear Regression **rất nhạy cảm với nhiễu** (sensitive to noise). Vi vậy trước khi thực hiện Linear Regressiong, các giá trị ngoại lai (outlier) cần phải được loại bỏ.

- Linear Regression **không biểu diễn được các mô hình phức tạp.**



Ví dụ 1:
Dự báo giá nhà

Bài toán dự đoán giá nhà.



Xây dựng mô hình học máy dự báo giá bán nhà dựa vào các thông tin liên quan

Bài toán dự đoán giá nhà.

- Tập dữ liệu bao gồm 545 mẫu, với các thuộc tính:

area: diện tích căn nhà (Số nguyên)

bedrooms: số phòng ngủ (Số nguyên)

bathrooms: số phòng tắm (Số nguyên)

mainroad: Có gần đường chính hay không (Yes|No)

gestroom: Có phòng khách hay không (Yes | No)

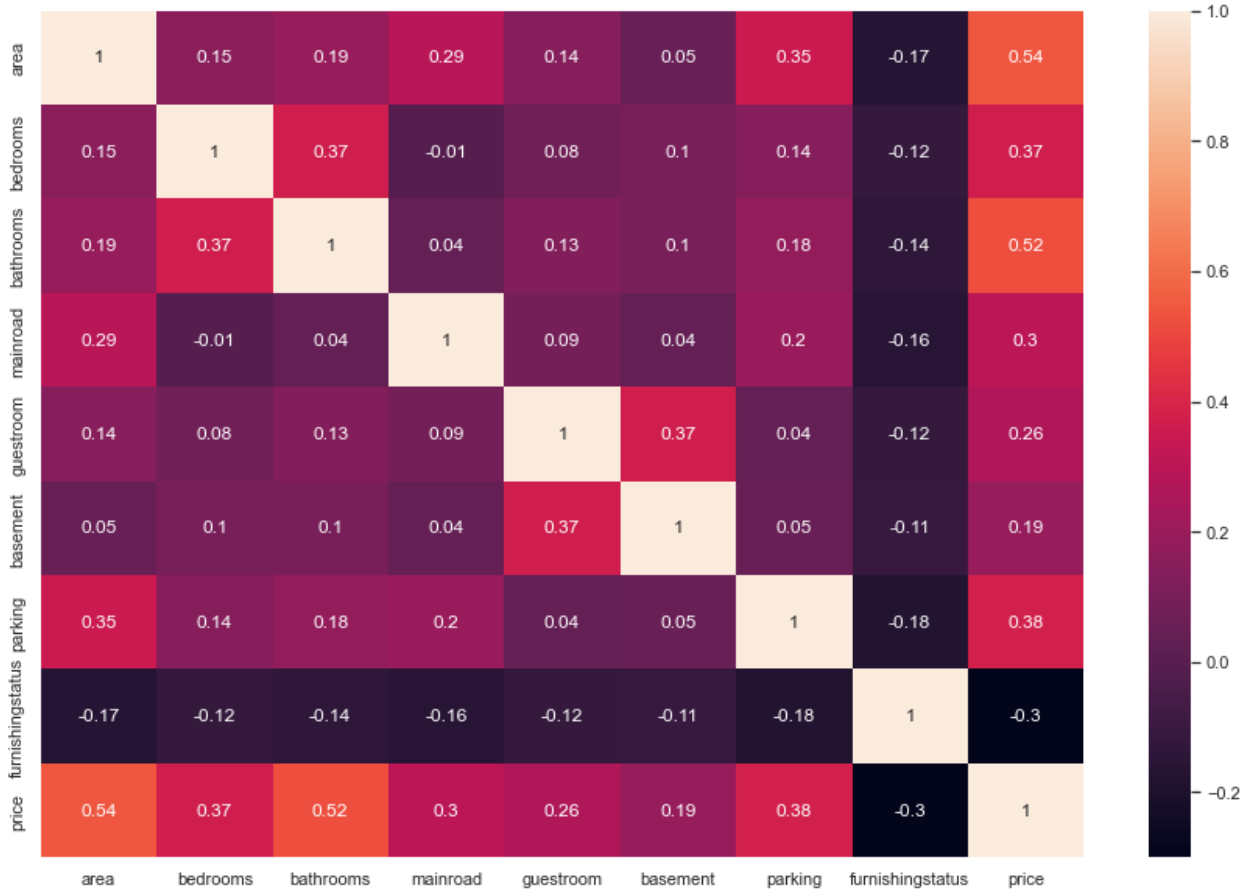
basement: Có tầng hầm hay không (Yes | No)

parking: số chỗ đỗ xe ôto (Số nguyên)

furnishingstatus: Tình trạng nội thất của căn nhà

- furnished: đầy đủ nội thất
- unfurnished: không có nội thất
- semi-furnished: được trang bị một phần nội thất

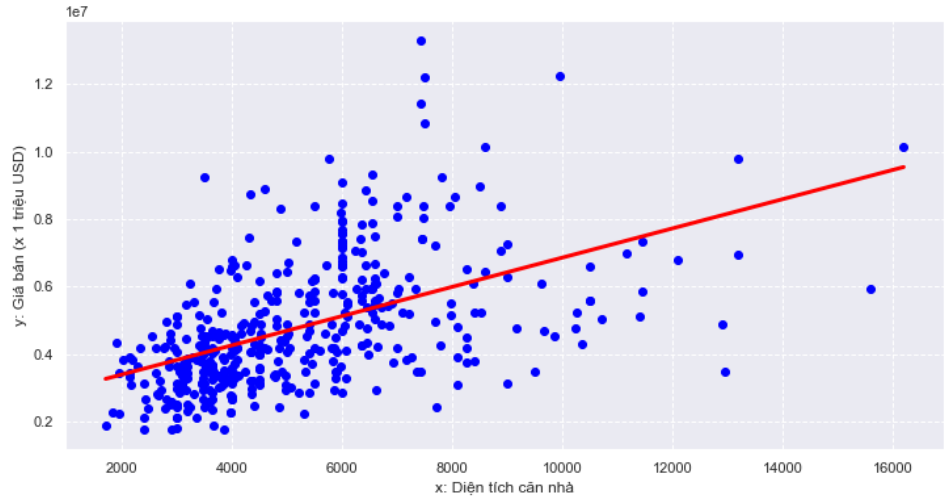
price: Giá bán ($)



Simple Linear Regression.

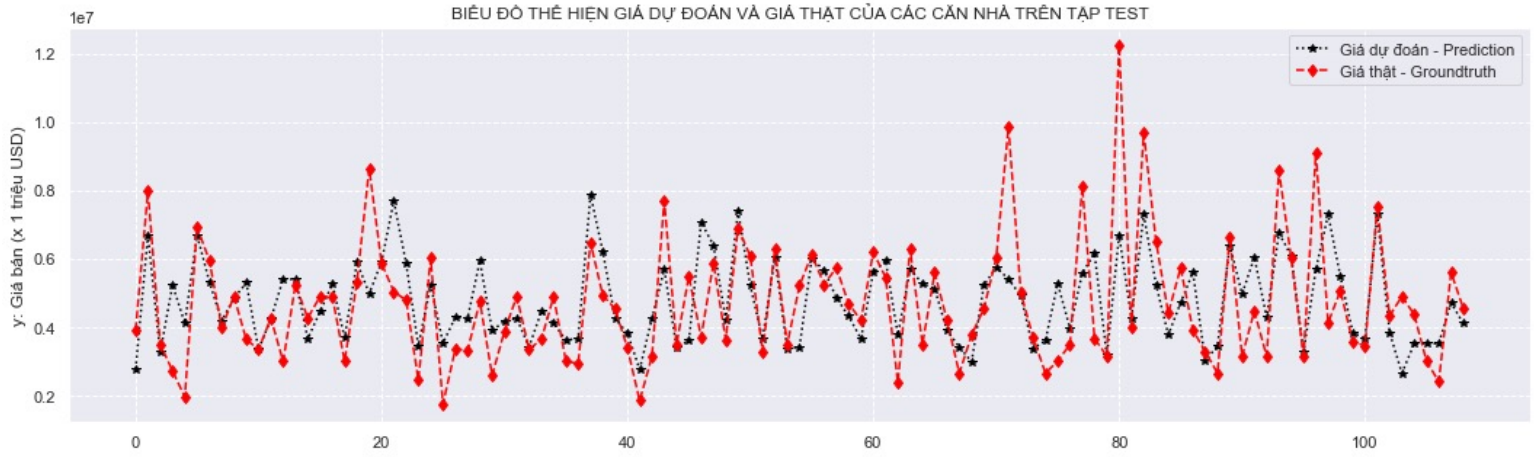
Dự đoán giá nhà với 1 biến độc lập – area (diện tích khu nhà)

	area	price
0	7420	13300000
1	8960	12250000
2	9960	12250000
3	7500	12215000
4	7420	11410000
5	7500	10850000
6	8580	10150000
7	16200	10150000
8	8100	9870000
9	5750	9800000



1.Sai số MAE = 1065153.0
2.Sai số MSE = 2132554869248.0
3.Sai số RMSE = 1460327.0

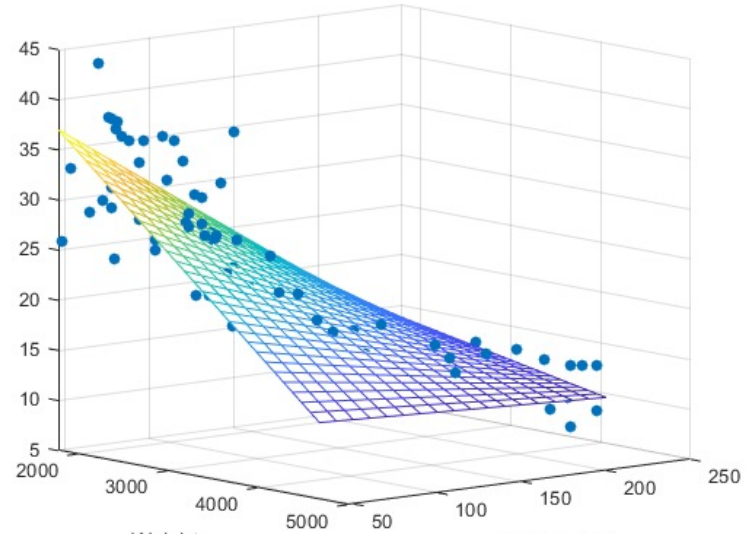
$\hat{y}_{MEDV} = f(x) = W_0 + W_1 * X_{RM} = 2530403.42 + 432.99 * X_{area}$



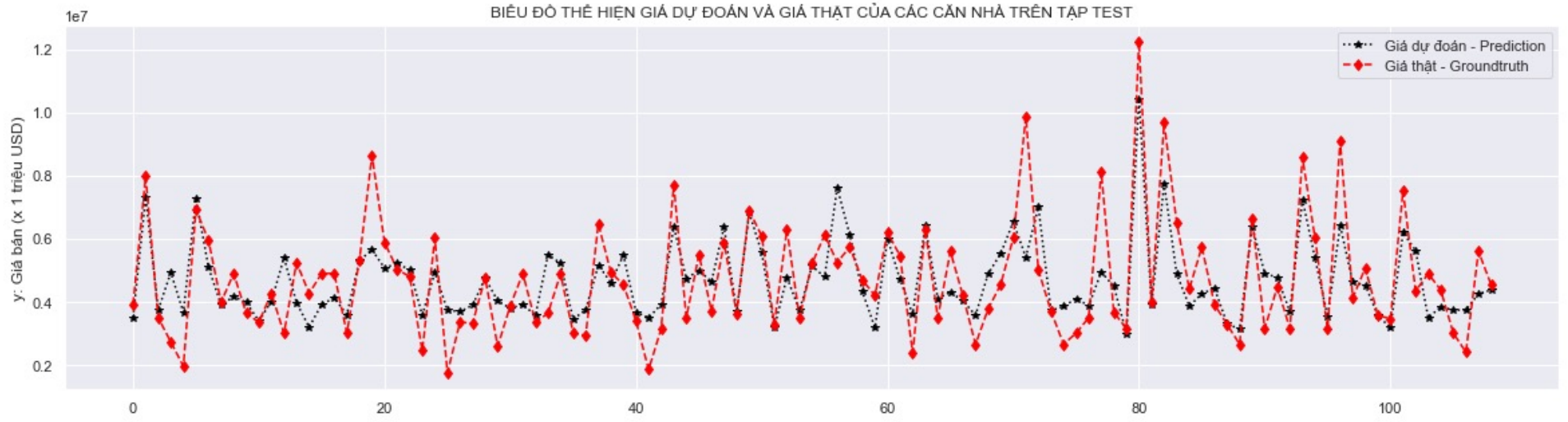
Multiple Linear Regression.

Sử dụng 2 thuộc tính (diện tích, số phòng tắm) để dự báo giá nhà:

	area	bathrooms	price
0	7420	2	13300000
1	8960	4	12250000
2	9960	2	12250000
3	7500	2	12215000
4	7420	1	11410000
5	7500	3	10850000
6	8580	3	10150000
7	16200	3	10150000
8	8100	1	9870000
9	5750	2	9800000




```
Độ chính xác của mô hình trên tập TEST:
----------------------------------------
1.Sai số MAE  =  874608.0
2.Sai số MSE =  1354595642165.0
3.Sai số RMSE =  1163871.0
```



Multiple Linear Regression.

Sử dụng tất cả 8 thuộc tính để dự báo giá nhà:

	area	bedrooms	bathrooms	mainroad	guestroom	basement	parking	furnishingstatus	price
0	7420	4	2	1	0	0	2	0	13300000
1	8960	4	4	1	0	0	3	0	12250000
2	9960	3	2	1	0	1	2	1	12250000
3	7500	4	2	1	0	1	3	0	12215000
4	7420	4	1	1	1	1	2	0	11410000
5	7500	3	3	1	0	1	2	1	10850000
6	8580	4	3	1	0	0	2	1	10150000
7	16200	5	3	1	0	0	0	2	10150000
8	8100	4	1	1	1	1	2	0	9870000
9	5750	3	2	1	1	0	1	2	9800000

	Coefficient
area	258.349
bedrooms	422944.324
bathrooms	1182870.919
mainroad	820199.779
guestroom	433374.993
basement	161753.541
parking	340674.754
furnishingstatus	-303174.544

1.Sai số MAE = 864732.0
2.Sai số MSE = 1141119782267.0
3.Sai số RMSE = 1068232.0

3. Đánh giá độ chính xác của mô hình hồi quy

Đánh giá mô hình hồi quy

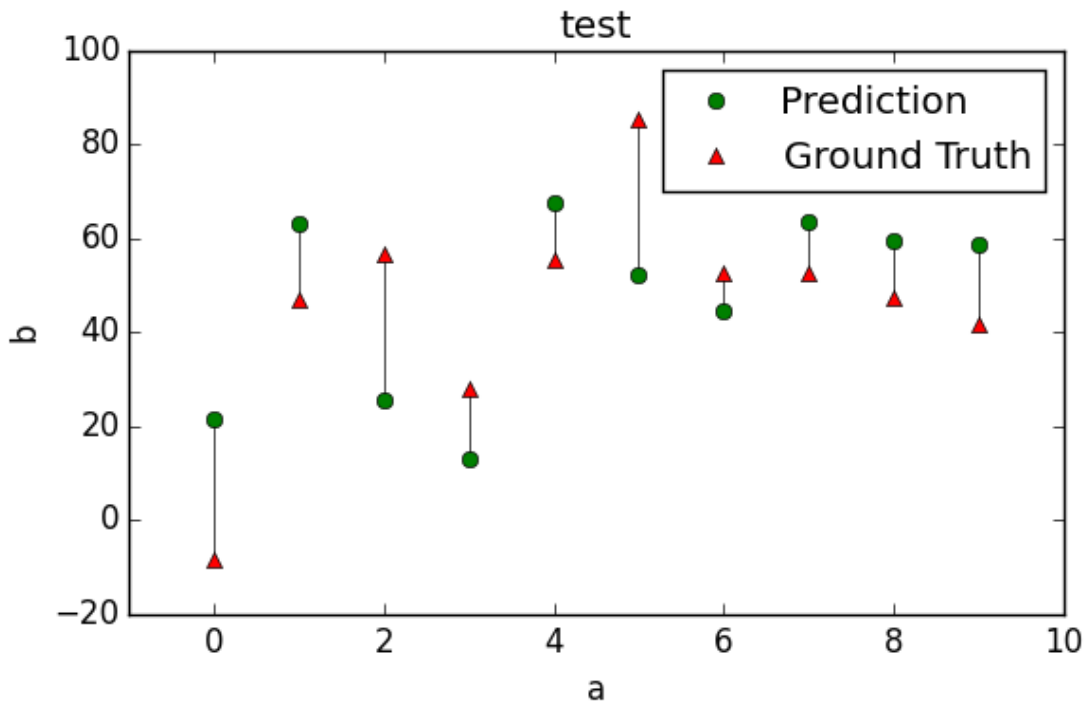
- Giả thiết: Có một mô hình học máy thực hiện việc dự đoán giá nhà tại một khu vực?
- Mô hình sau khi được huấn luyện với dữ liệu Training, thực hiện kiểm thử mô hình trên tập dữ liệu Test với số lượng 100 mẫu.

Evaluating Machine Learning Models

y_predict	y_groundtruth
22 890	23 432
19 120	18 850
9 590	10 500
20 231	22 567
7 498	5 235
13 675	11 563
22 453	25 005
24 645	19 214
30 654	27 087
5 643	8 675
14 087	13 675
8 000	7 465
25 986	29 875

Đánh giá mô hình hồi quy

- Các chỉ số cơ bản để đánh giá độ chính xác của mô hình hồi quy:



Regression

- *MAE*

(mean abs. error)

- *MSE*

(mean sq. error)

- *RMSE*

(Root mean sq.error)

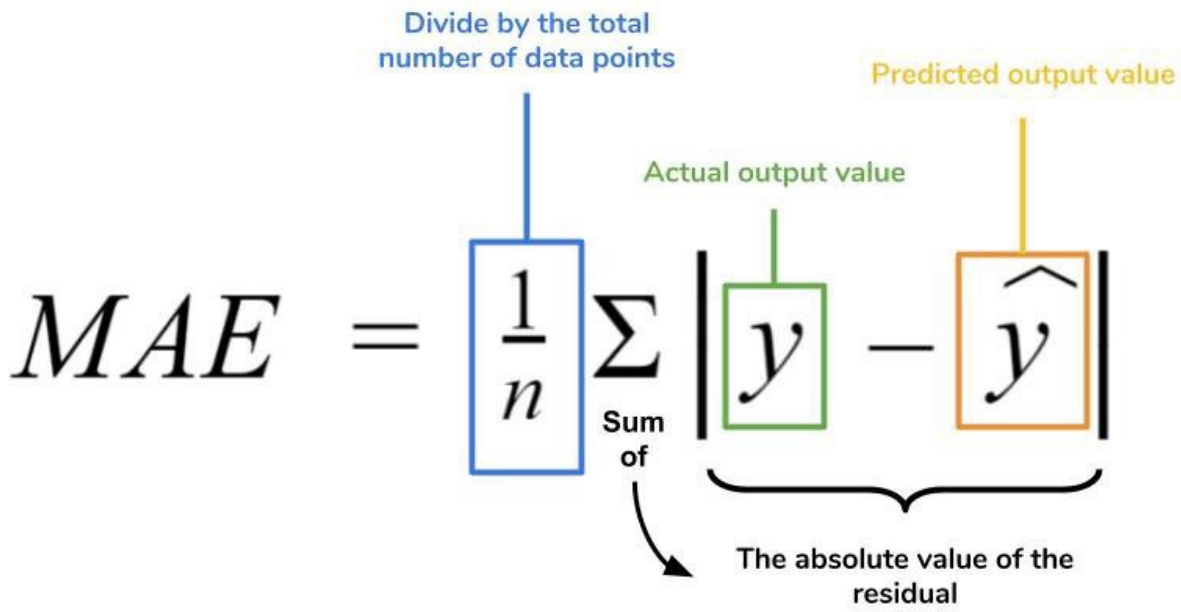
- *RMSLE*

(Root mean sq.error log error)

- R^2 *and Adjusted* R^2

1. Sai số MAE

- Sai số tuyệt đối trung bình (MAE – Mean Absolute Error) nằm trong khoảng (0, +∞). MAE biểu thị biên độ trung bình của sai số mô hình nhưng không nói lên xu hướng lệch của giá trị dự đoán (predicted) và giá trị thực (Actual). Khi MAE = 0, các giá trị dự đoán hoàn toàn trùng khớp với các giá trị thực, khi đó mô hình được xem là "lý tưởng"



2. Sai số MSE

- **Sai số bình phương trung bình (MSE)** nằm trong khoảng (0, +∞), MSE phản ánh mức độ dao động giữa giá trị dự đoán với giá trị thực.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

3. Sai số RMSE

- **Sai số bình phương trung bình quân phương (RMSE)** là một trong những đại lượng cơ bản và thường được sử dụng phổ biến trong đánh giá độ tin cậy của mô hình hồi quy. Người ta thường hay sử dụng RMSE biểu thị độ lớn trung bình của sai số. Đặc biệt RMSE rất nhạy với những giá trị sai số lớn. Giống như MAE, RMSE không chỉ ra độ lệch giữa giá trị dự báo và giá trị thực. Giá trị của RMSE nằm trong khoảng (0,+ ∞)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

4. Hệ số R^2

R2: Đánh giá tỷ lệ giải thích của mô hình ước lượng, hệ số này nằm giữa 0 và 1, càng gần 1 tỷ lệ giải thích được của mô hình càng tốt.

- Giá trị R bình phương dao động từ 0 đến 1. R bình phương càng gần 1 thì mô hình đã xây dựng càng phù hợp với bộ dữ liệu dùng chạy hồi quy. R bình phương càng gần 0 thì mô hình đã xây dựng càng kém phù hợp với bộ dữ liệu dùng chạy hồi quy. Trường hợp đặt biệt, phương trình hồi quy đơn biến (chỉ có 1 biến độc lập) thì R2 chính là bình phương của hệ số tương quan r giữa hai biến đó.

$$\textit{Coefficient of Determination} \rightarrow \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

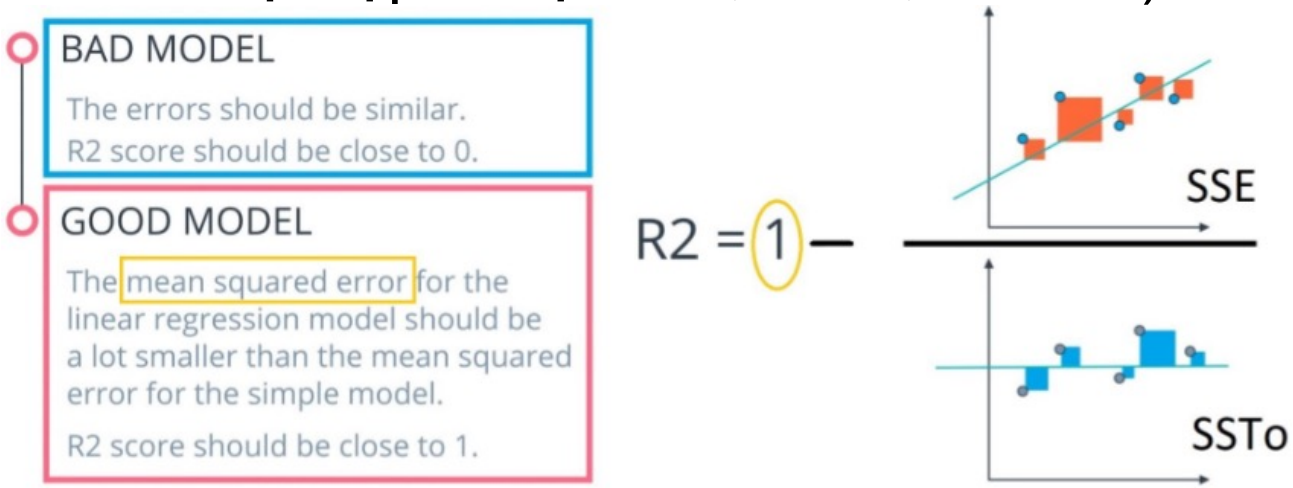
$$\textit{Sum of Squares Total} \rightarrow \quad SST = \sum (y - \bar{y})^2$$

$$\textit{Sum of Squares Regression} \rightarrow \quad SSR = \sum (y' - \bar{y'})^2$$

$$\textit{Sum of Squares Error} \rightarrow \quad SSE = \sum (y - y')^2$$

4. Hệ số R^2

Ý nghĩa R bình phương: Giả sử R bình phương là 0.60, thì mô hình hồi quy tuyến tính này phù hợp với tập dữ liệu ở mức 60%. Nói cách khác, 60% biến thiên của biến phụ thuộc được giải thích bởi các biến độc lập.(còn 40% còn lại ở đâu, dĩ nhiên là do sai số đo lường, do cách thu thập dữ liệu, do có thể có biến độc lập khác giải thích cho biến phụ thuộc mà chưa được được vào mô hình nghiên cứu…vv). Thông thường, ngưỡng của R2 phải trên 50%, vì như thế mô hình mới phù hợp. Tuy nhiên tùy vào dạng nghiên cứu, như các mô hình về tài chính, không phải tất cả các hệ số R2 đều bắc buộc phải thỏa mãn lớn hơn 50%.(do rất khó dể dự đoán giá vàng, giá cổ phiếu mà chỉ đơn thuần dựa vào các biến độc lập ví dụ GDP, ROA,ROE….)



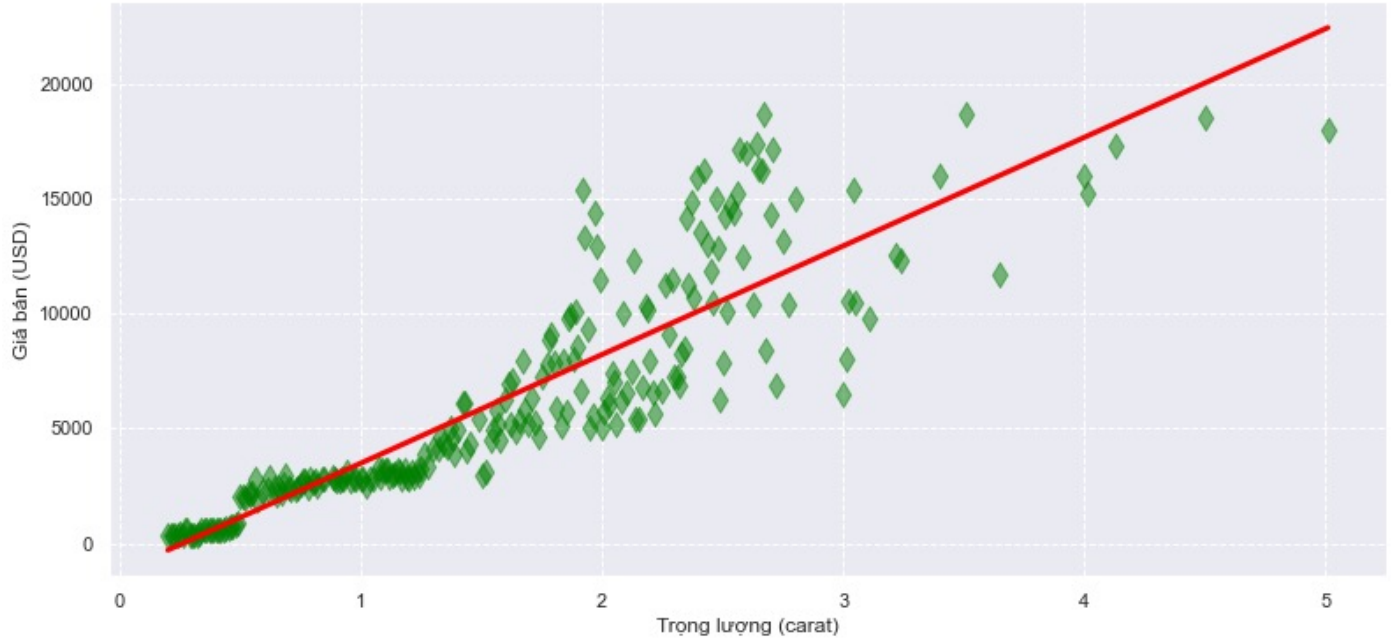
THỰC HÀNH

Yêu cầu

Tập dữ liệu Data_Diamonds.xlxs lưu trữ dữ liệu của 273 viên kim cương. Xây dựng mô hình hồi quy tuyến tính đơn giản sử dụng biến trọng lượng (carat) để dự doán giá bán kim cương (price.

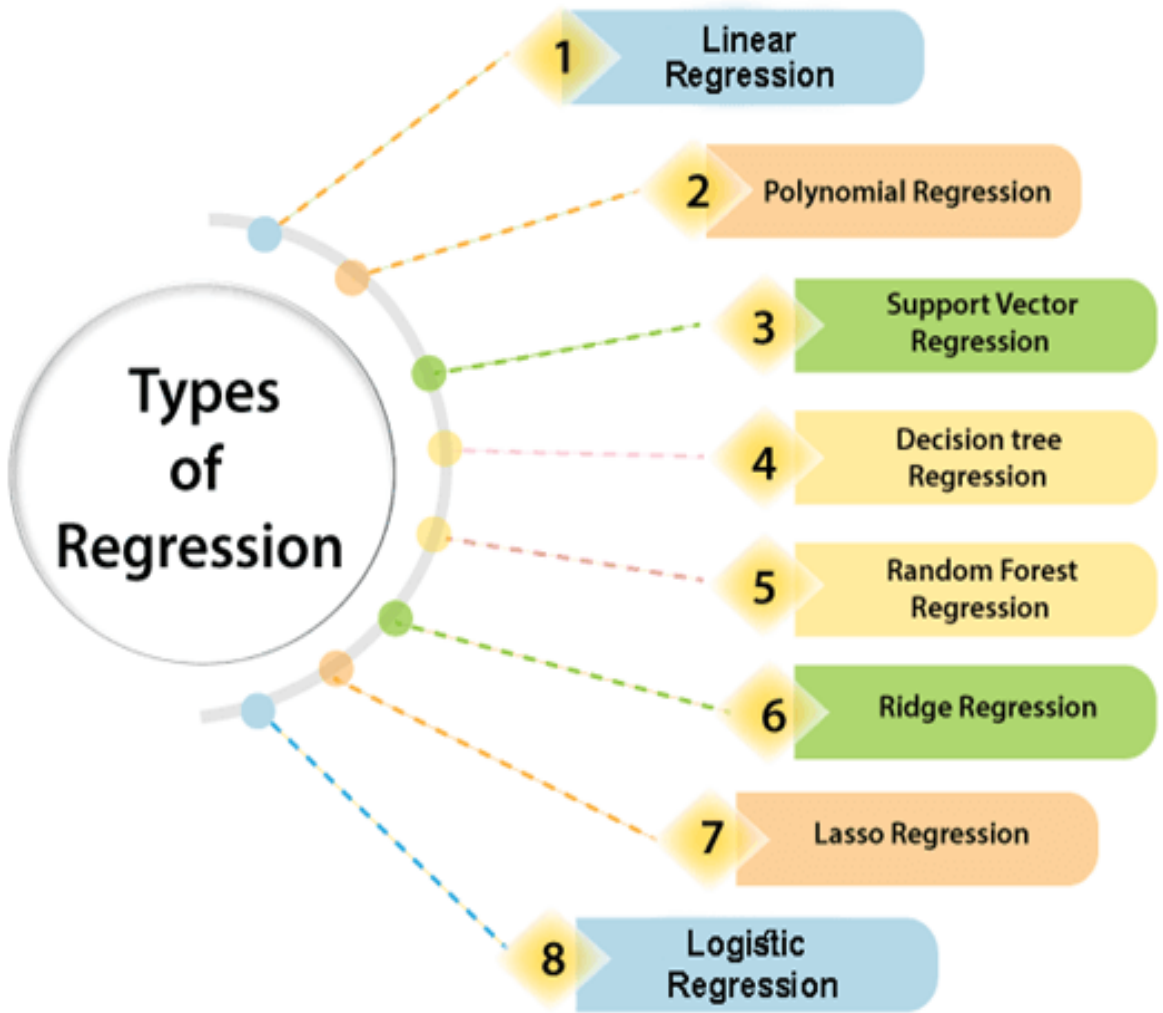
1. Sử dụng 85% dữ liệu viên kim cương để huấn luyện mô hình với tham số mặc định; Hiển thị biểu đồ các viên kim cương tập train và đường hồi quy.
2. Sử dụng mô hình trên 15% của tập Test, Xác định các sai số MAE, MSE, RMSE của mô hình; hiển thị biểu đồ thể hiện giá thật và giá dự đoán của tập test
3. Sử dụng mô hình dự đoán viên kim cương có trọng lượng 2.88 carat có giá bán bao nhiêu?

carat	cut	color	clarity	depth	table	price
0.21	Premium	E	SI1	59.8	61	326
0.23	Ideal	E	SI2	61.5	55	326
0.29	Premium	I	VS2	62.4	58	334
0.31	Good	J	SI2	63.3	58	335
0.24	Very Good	J	VVS2	62.8	57	336
0.22	Fair	E	VS2	65.1	61	337
0.26	Very Good	H	SI1	61.9	55	337
0.3	Good	J	SI1	64	55	339
0.2	Premium	E	SI2	60.2	62	345
0.32	Premium	E	I1	60.9	58	345
0.33	Ideal	I	SI2	61.8	55	403
0.25	Very Good	E	VS2	63.3	60	404
0.35	Ideal	I	VS1	60.9	57	552
0.42	Premium	I	SI2	61.5	59	552



4. Một số mô hình hồi quy khác (Sinh viên tìm hiểu thêm)

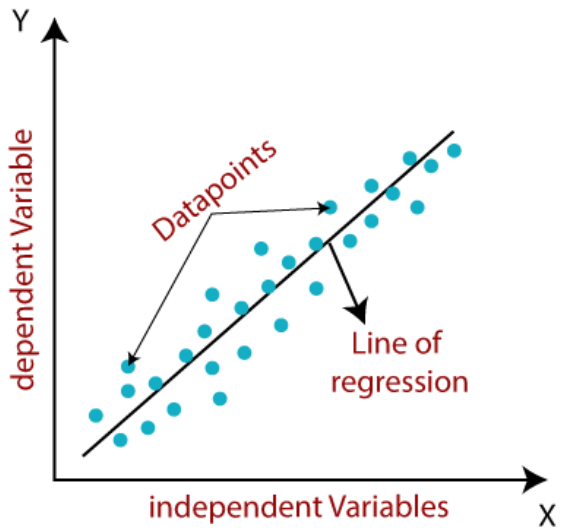
Một sô mô hình hồi quy

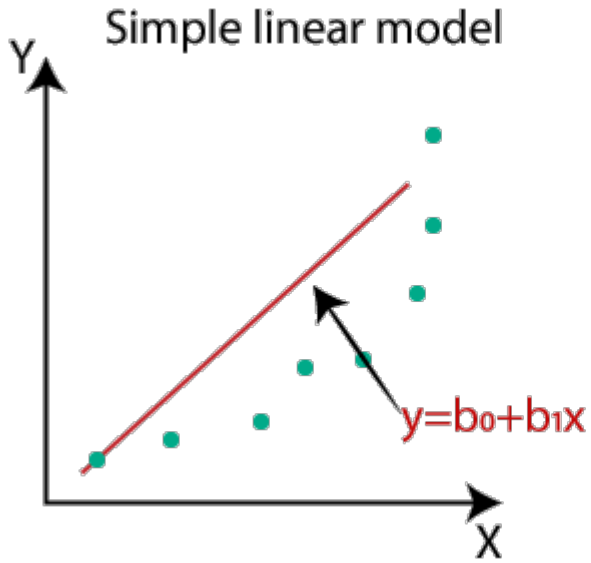


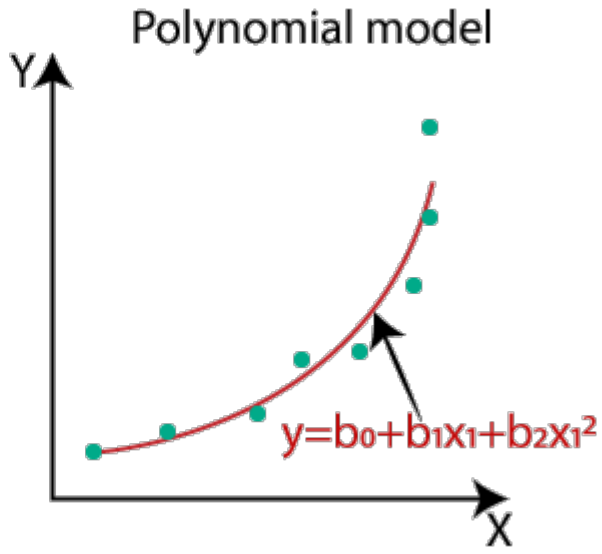
4.1 Hồi quy đa thức (Polynomial Regression)

Hồi quy đa thức

Trong trường hợp dữ liệu không tuyến tính việc áp dụng mô hình tuyến tính sẽ không hiệu quả tỷ lệ lỗi cao, độ chính xác giảm.







Hồi quy đa thức bậc n của biến độc lập x_1

Polynomial Linear Regression

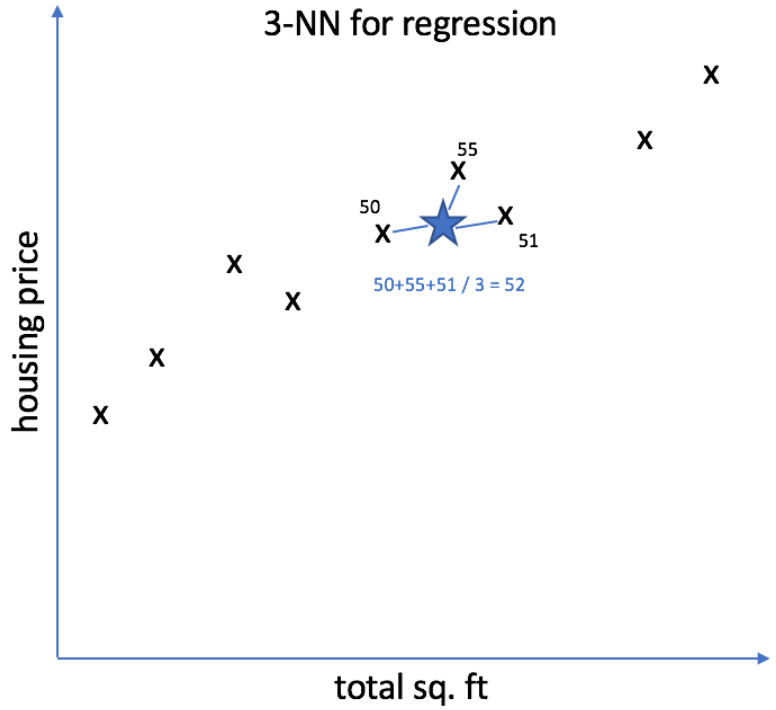
$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

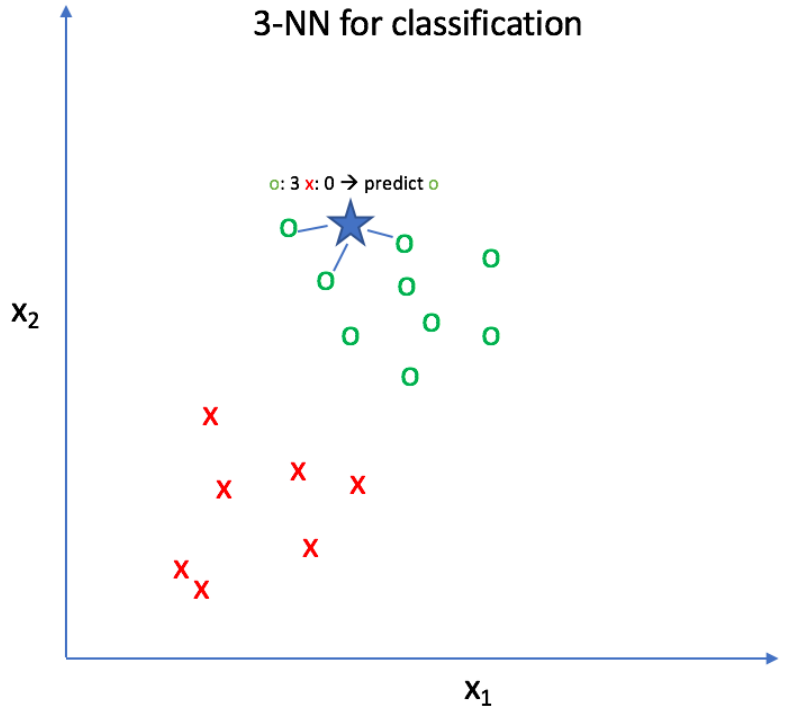
4.2 KNN cho bài toán Hồi quy

KNN

Tương tự như đối với bài toán phân lớp. Xác định những điểm dữ liệu gần nhất với điểm dữ liệu mới.

Nhãn của điểm dữ liệu mới được là nhãn của điểm dữ liệu đã biết gần nhất (K=1) hoặc trung bình có trọng số của những điểm gần nhất.

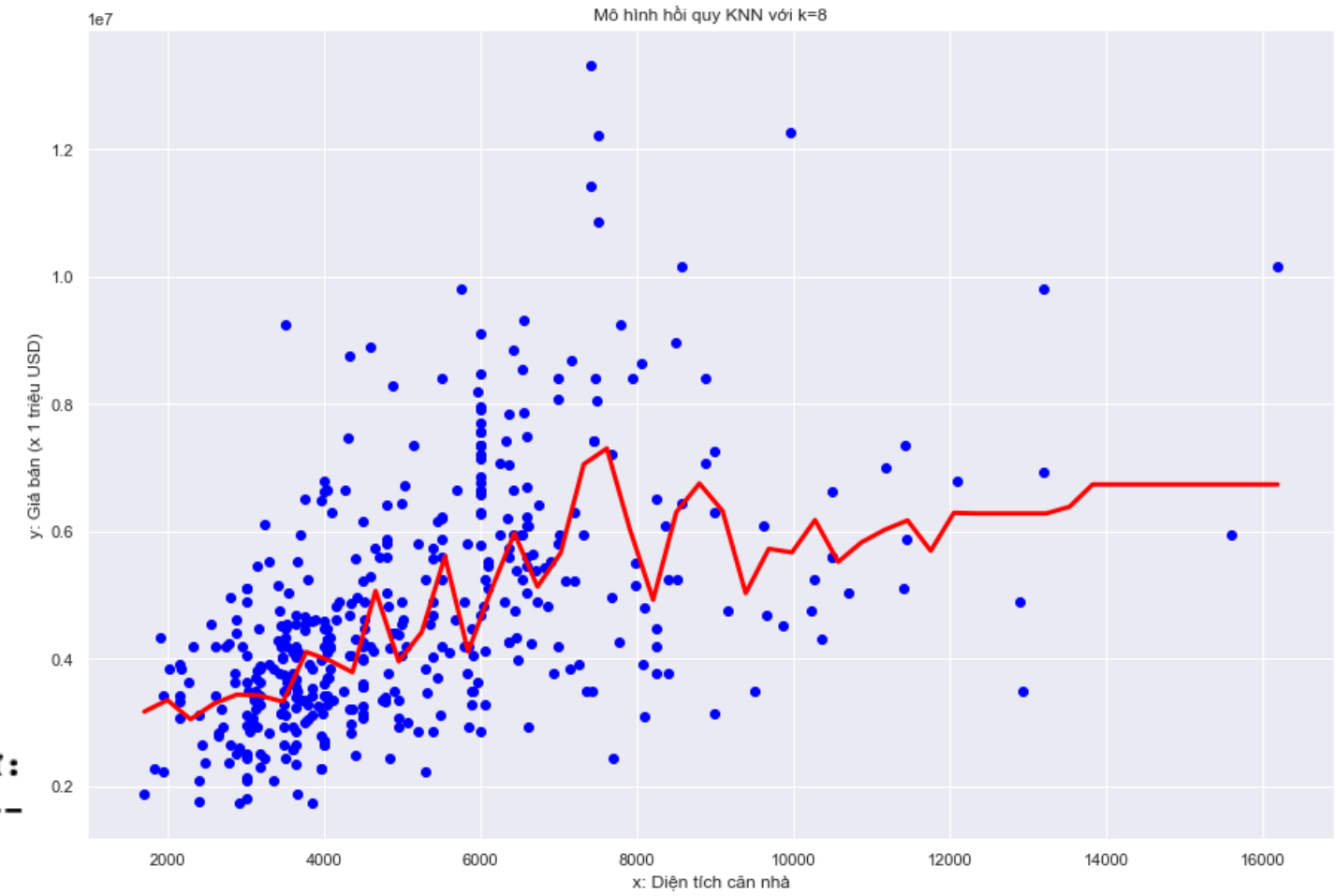




KNN

Dự đoán giá nhà với 1 biến độc lập – RM (số phòng trung bình của căn nhà)

	area	price
0	7420	13300000
1	8960	12250000
2	9960	12250000
3	7500	12215000
4	7420	11410000
5	7500	10850000
6	8580	10150000
7	16200	10150000
8	8100	9870000
9	5750	9800000




```
Độ chính xác của mô hình trên tập kiểm thử:
-------------------------------------------
1. Sai số MAE 864732.1743119266
2. Sai số RMSE   1423888.4591579423
```


HANOI UNIVERSITY OF MINING AND GEOLOGY
KHOA
CÔNG NGHỆ THÔNG TIN
FACULTY OF INFORMATION TECHNOLOGY

Thank you!