REGRESSION WITH CRAB AGE DATASET

Report Challenge

Lê Việt Hưng
Phạm Ngọc Hiếu

Content

1 Data descrciption

2 Data Visualization

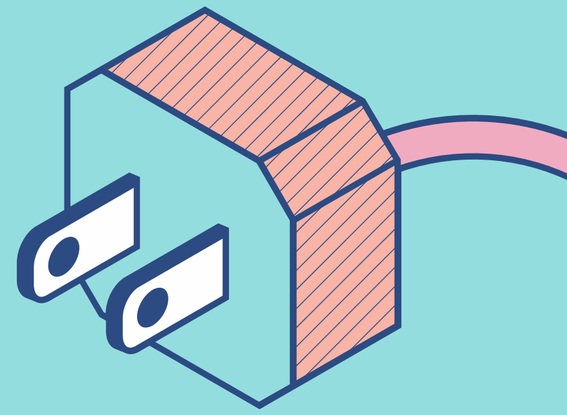
3 Pre-processing

4 Feature Engineering

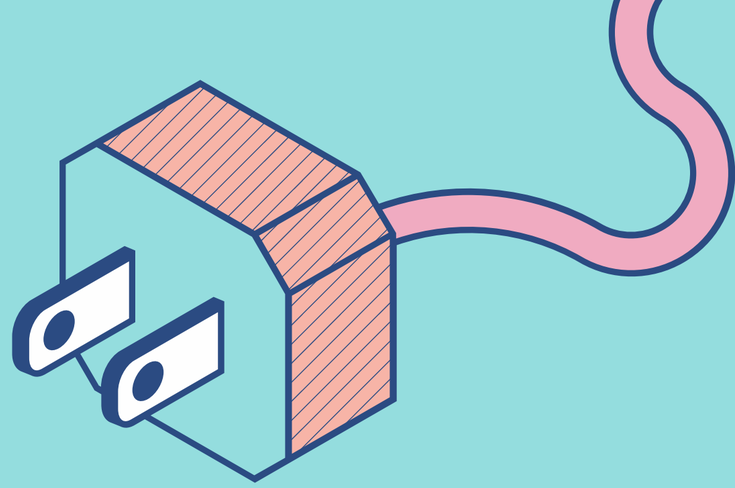
5 Model

1.Data Description

Dataset gồm 2 file csv: train.csv và test.csv
train.csv là dữ liệu dạng bảng chứa 9 columns trong có 8 columns là features và 1 column là target
test.csv là dữ liệu dạng bảng chứa 9 columns trong có 8 columns là features



1.Data Description

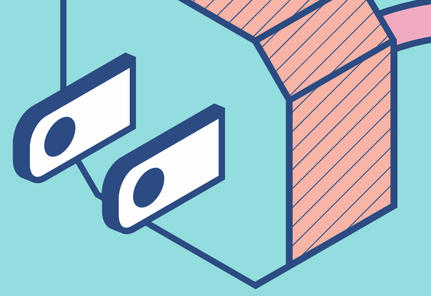


Ta sử dụng thư viện pandas để đọc dữ liệu (train.csv) đầu vào:

	id	Sex	Length	Diameter	Height	Weight	Shucked Weight	Viscera Weight	Shell Weight	Age
0	0	I	1.5250	1.1750	0.3750	28.973189	12.728926	6.647958	8.348928	9
1	1	I	1.1000	0.8250	0.2750	10.418441	4.521745	2.324659	3.401940	8
2	2	M	1.3875	1.1125	0.3750	24.777463	11.339800	5.556502	6.662133	9
3	3	F	1.7000	1.4125	0.5000	50.660556	20.354941	10.991839	14.996885	11
4	4	I	1.2500	1.0125	0.3375	23.289114	11.977664	4.507570	5.953395	8

- Features: Sex,Length,Diameter,Height,Weight,Shucked Weight,Viscera Weight, Shell Weight
- Target: Age

1.Data Description

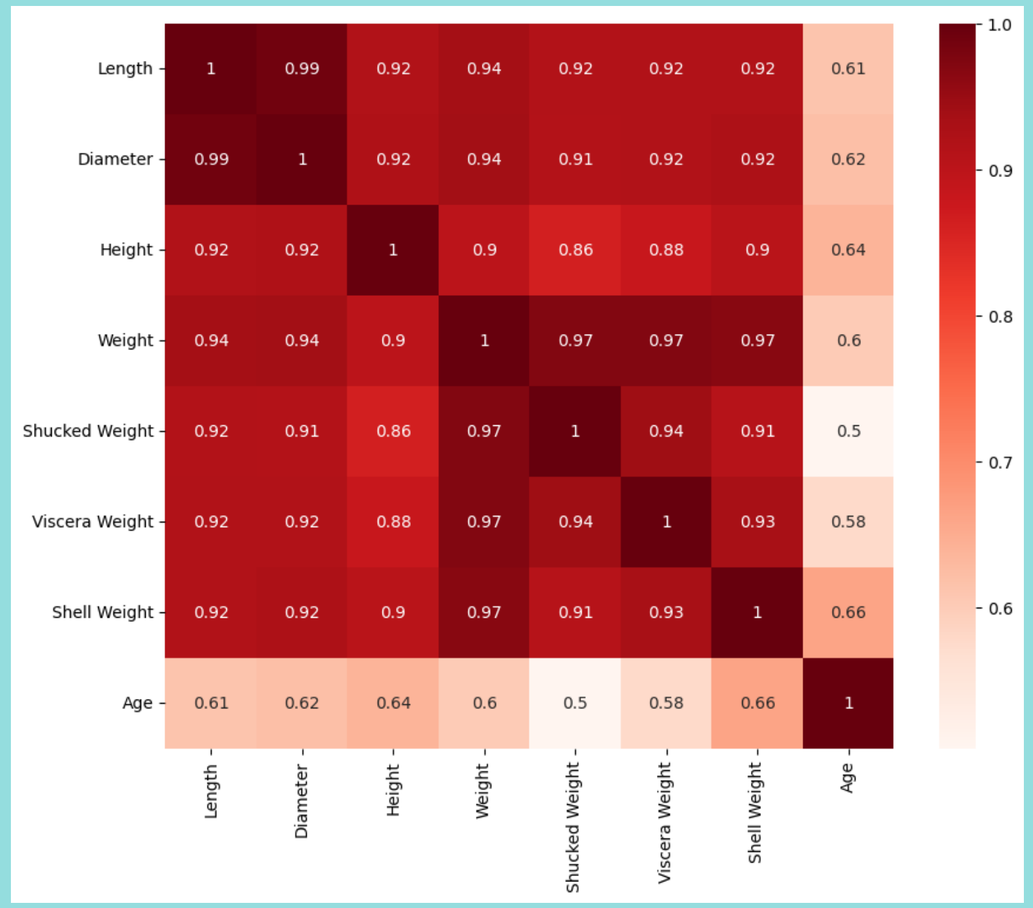


Giải thích các đặc trưng :

1. Sex: Giới tính của con cua
2. Length: Độ dài của con cua từ đầu đến đuôi của con cua
3. Diameter: Đường kính của con cua
4. Height: Chiều cao của con cua
5. Weight: Trọng lượng của con cua
6. Shucked Weight: Trọng lượng thịt cua sau khi đã tách vỏ
7. Viscera Weight: Trọng lượng của phần ruột của con cua
8. Shell Weight: Trọng lượng vỏ của con cua

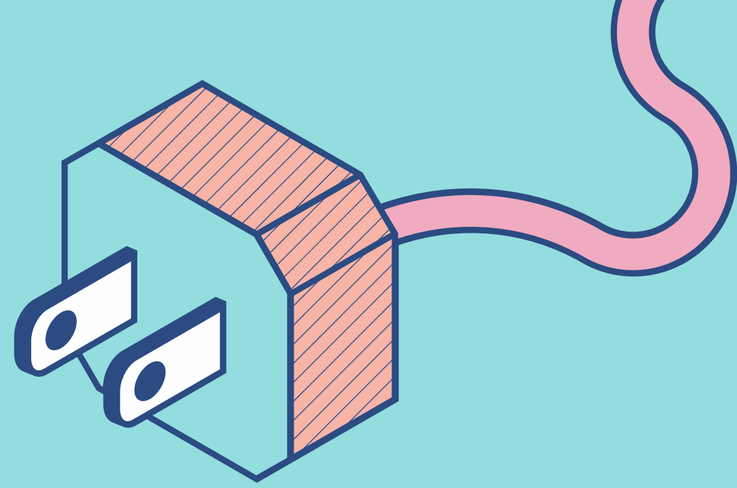
2. Data Visualization

Matrix Correlation:



Các Features có độ tương quan -> cần tạo ra các đặc trưng mới có độ tương quan với nhau thấp hơn

3. Pre-processing



1.Có một số điểm dữ liệu tính năng "Height" có giá trị bằng 0 => Vô lý

=>Cách giải quyết: Sử dụng RandomForestRegressor, được huấn luyện trên tập dữ liệu train để dự doán các giá trị "Height" thay cho "Height" = 0

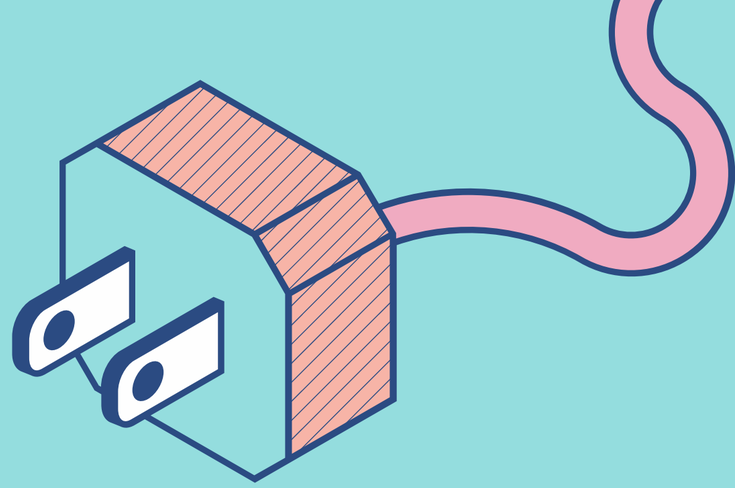
2.Sử dụng kĩ thuật "Encoding" để chuyển giá trị "Sex" về số, bao gồm "Sex_M" và "Sex_I"

"Sex_M" bằng 1 nghĩa là giới tính nam, bằng 0 nghĩa là giới tính nữ

"Sex_I" bằng 1 nghĩa là giới tính không xác định

Không sử dụng encode kiểu "Ordinal Encoding" vì như vậy nếu giới tính không xác định sẽ gần giới tính nữ hơn là giới tính nam => không đúng

4. Feature Engineering



Sử dụng Feature Engineering để tạo ra các đặc trưng mới có độ tương quan với nhau thấp hơn

Viscera Ratio:

$$\text{Viscera Ratio} = \frac{\text{Viscera Weight}}{\text{Weight}}$$

Shell Ratio:

$$\text{Shell Ratio} = \frac{\text{Shell Weight}}{\text{Weight}}$$

Surface Area:

$$\text{Surface Area} = 2 \times (\text{Length} \times \text{Diameter} + \text{Length} \times \text{Height} + \text{Diameter} \times \text{Height})$$

Volume:

$$\text{Volume} = \text{Length} \times \text{Diameter} \times \text{Height}$$

Density:

$$\text{Density} = \frac{\text{Weight}}{\text{Volume}}$$

Shell-to-Body Ratio:

$$\text{Shell-to-Body Ratio} = \frac{\text{Shell Weight}}{\text{Weight} + \text{Shell Weight}}$$

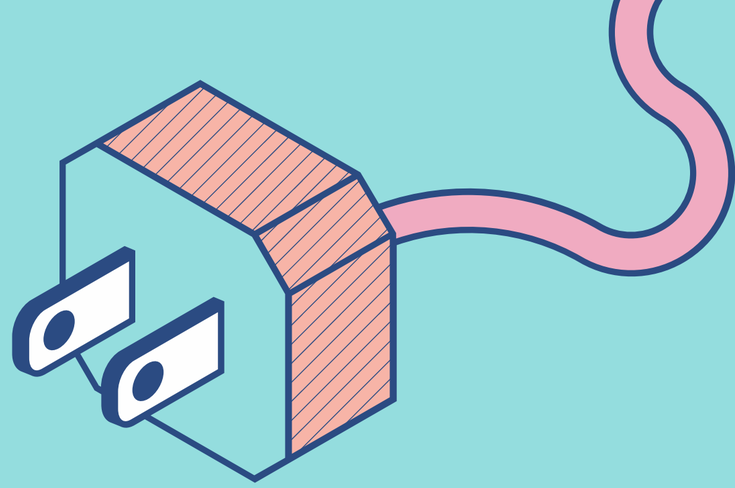
Meat Yield:

$$\text{Meat Yield} = \frac{\text{Shucked Weight}}{\text{Weight} + \text{Shell Weight}}$$

Body Condition Index:

$$\text{Body Condition Index} = \sqrt{\text{Length} \times \text{Weight} \times \text{Shucked Weight}}$$

4. Feature Engineering



Len-to-Diam:

$$\text{Len-to-Diam} = \frac{\text{Length}}{\text{Diameter}}$$

wieght-to-Viswieght:

$$\text{wieght-to-Viswieght} = \frac{\text{Weight}}{\text{Viscera Weight}}$$

wieght-to-Shellwieght:

$$\text{wieght-to-Shellwieght} = \frac{\text{Weight}}{\text{Shell Weight}}$$

wieght-to-Shckwieght:

$$\text{wieght-to-Shckwieght} = \frac{\text{Weight}}{\text{Shucked Weight}}$$

Weight_wo_Viscera:

$$\text{Weight\_wo\_Viscera} = \text{Shucked Weight} - \text{Viscera Weight}$$

Length^2:

$$\text{Length}^2 = \text{Length}^2$$

Diameter^2:

$$\text{Diameter}^2 = \text{Diameter}^2$$

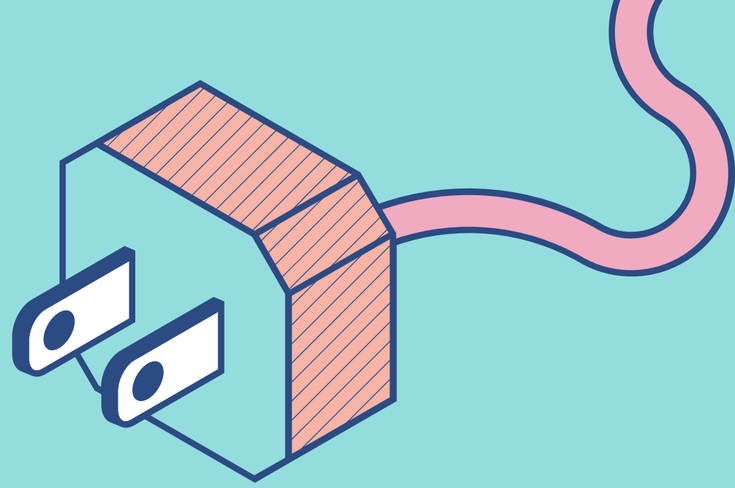
Log Length:

$$\text{Log Length} = \log(\text{Length} + 1)$$

Log Diameter:

$$\text{Log Diameter} = \log(\text{Diameter} + 1)$$

4. Feature Engineering



Log Height:

$$\text{Log Height} = \log(\text{Height} + 1)$$

Weight^2:

$$\text{Weight}^2 = \text{Weight}^2$$

Shucked Weight^2:

$$\text{Shucked Weight}^2 = \text{Shucked Weight}^2$$

Viscera Weight^2:

$$\text{Viscera Weight}^2 = \text{Viscera Weight}^2$$

Shell Weight^2:

$$\text{Shell Weight}^2 = \text{Shell Weight}^2$$

Meat Ratio:

$$\text{Meat Ratio} = \frac{\text{Shucked Weight}}{\text{Weight}}$$

Weight-to-Volume Ratio:

$$\text{Weight-to-Volume Ratio} = \frac{\text{Weight}}{\text{Volume}}$$

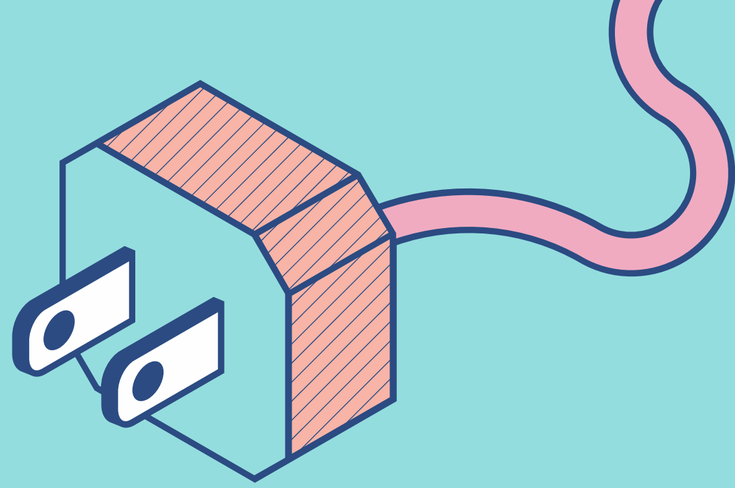
Weight-to-Length Ratio:

$$\text{Weight-to-Length Ratio} = \frac{\text{Weight}}{\text{Length}}$$

Weight-to-Height Ratio:

$$\text{Weight-to-Height Ratio} = \frac{\text{Weight}}{\text{Height}}$$

4. Feature Engineering



Shucked Weight per Length:

$$\text{Shucked Weight per Length} = \frac{\text{Shucked Weight}}{\text{Length}}$$

Viscera Weight per Length:

$$\text{Viscera Weight per Length} = \frac{\text{Viscera Weight}}{\text{Length}}$$

Shell Weight per Length:

$$\text{Shell Weight per Length} = \frac{\text{Shell Weight}}{\text{Length}}$$

Length-Diameter Difference:

$$\text{Length-Diameter Difference} = |\text{Length} - \text{Diameter}|$$

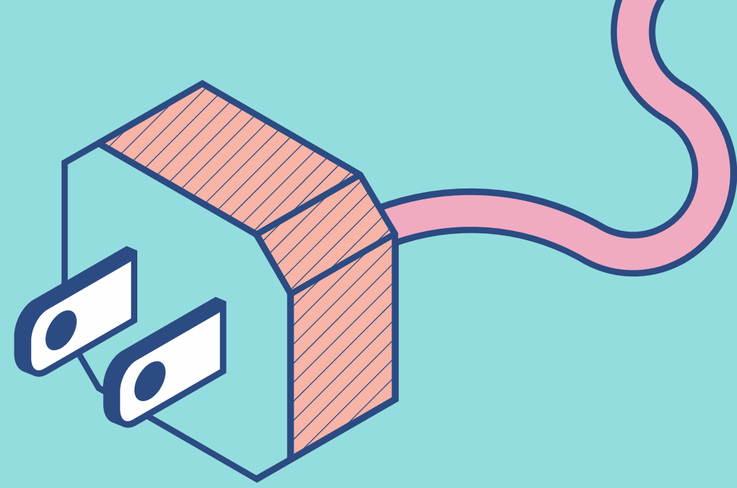
Weight-Diameter Ratio:

$$\text{Weight-Diameter Ratio} = \frac{\text{Weight}}{\text{Diameter}}$$

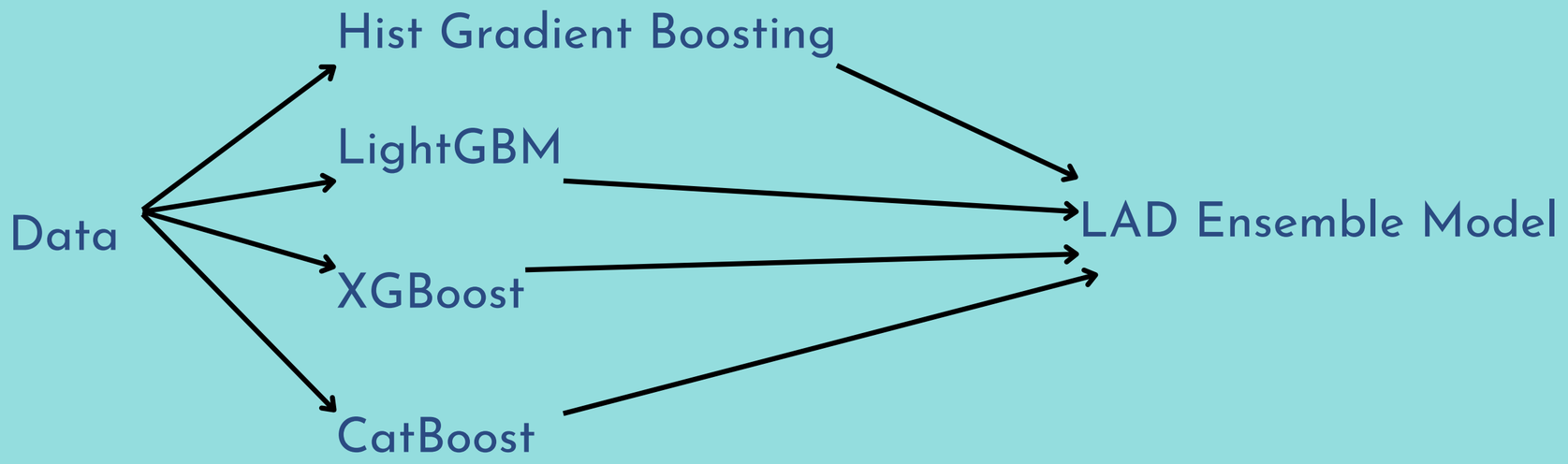
Weight to Length ^2:

$$\text{Weight-to-Length Ratio} = \frac{\text{Weight}}{\text{Length}^2}$$

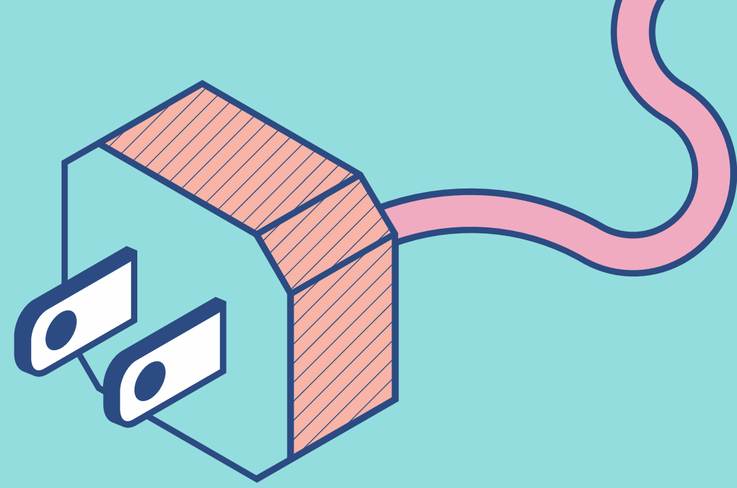
5. Model



Sử dụng Stacking Ensemble Learning:



5. Model



Đối với các base-model, thực hiện tìm kiếm theo lưới để chọn ra các tham số có giá trị tốt nhất

Vì tập dữ liệu không lớn nên sẽ sử dụng k-fold cross-validation để đánh giá và chọn ra các tham số và tập tính năng tốt nhất cho từng model

Sau đó, sử dụng các model với tính năng và tham số tốt nhất để làm đầu vào cho meta-model(LAD Regression) đưa ra dự đoán cuối cùng

Kết quả tốt nhất đạt được là 1.33568 đối với private score và 1.33627 đối với public score

Thank you