Playing with Machine Learning in the Kaggle Playground

- Final project -

2조

강민기 ㅇㅇㅇ
ㅇㅇㅇ ㅇㅇㅇ
ㅇㅇㅇ

Advisor : ㅇㅇㅇ & ㅇㅇㅇ 멘토님

24.04.26

목차

01 팀 소개

- 팀 소개
- Workflow

02 분류(Classification)

- 대회 소개 및 선정 배경
- 데이터 설명
- 데이터 전처리
- EDA
- 모델링
- 대회 제출 결과
- 인사이트

03 회귀(Regression)

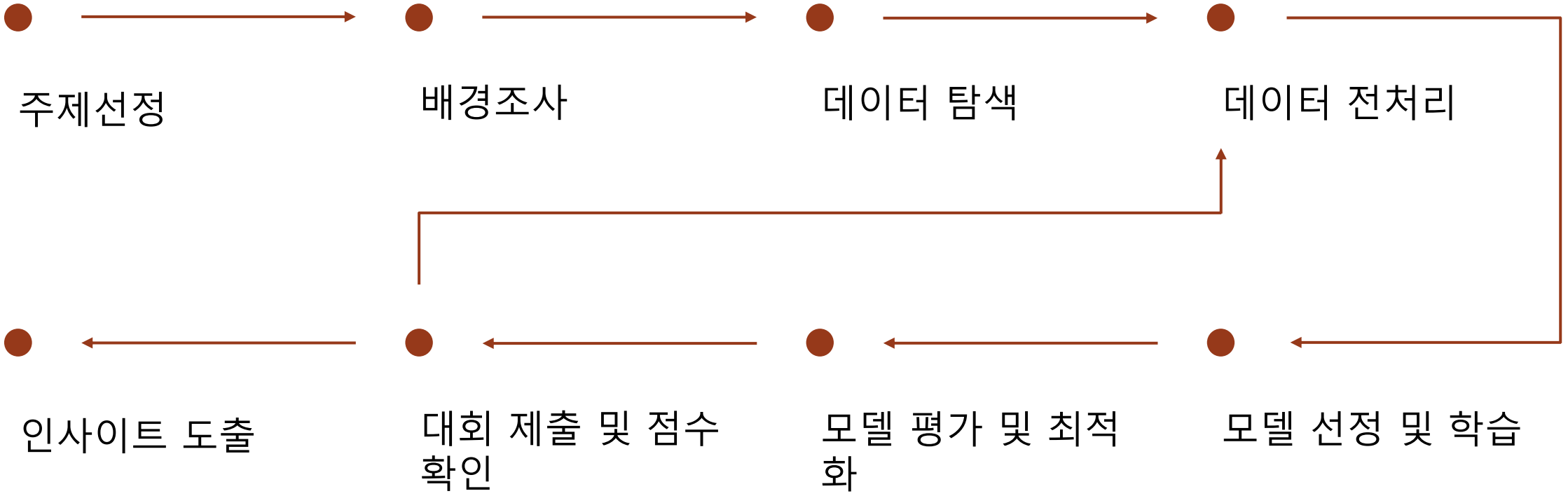
- 대회 소개 및 선정 배경
- 데이터 설명
- 데이터 전처리
- EDA
- 모델링
- 대회 제출 결과
- 인사이트

04 Q&A

- 질의응답

팀소개

팀소개 및 WorkFlow



분류(Classification)

대회 소개

01 대회 제목

- Binary Classification with a Bank Churn Dataset

02 모델 종류

- **분류**(Binary Classification)

03 대회 개요

- 은행 고객의 **이탈 여부 예측**

04 평가 지표

- **ROC 곡선 아래 영역**을 측정하여 평가 지표로 활용

분류 파트 삭제

회귀(Regression)

대회 소개

01 대회 제목

- **Regression with an Abalone Dataset**

02 평가 지표

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

(Root Mean Squared Logarithmic Error)

03 대회 개요

- 전복의 외적인 특성을 이용하여 전복의 나이테 수를 예측

04 모델 종류

- 회귀(Regression)

대회 선정 배경

01 주제 선정 이유

- 외부적인 측정값으로부터 내부적 특성을 유추하는 모델은 다양한 분야에서 자주 등장
- 배경지식에 기반하여 **새로운 피처를 생산해 내는 능력 향상**에 도움 될 것이라고 판단

02 주제를 통해 세운 가설

- 전복도 사람과 마찬가지로 나이에 따라 **BMI**가 다를 것이라 예상
- 전복의 **성장 속도**는 (전복의 고기 및 내장 무게)/(전복 전체 무게)와 관련
- 전복 크기는 길이 및 너비를 각각 고려하는 것보다 **면적(길이*너비)**이 더 관련

03 전복 나이를 예측한 활용 방안 제시

- 나이에 따라 전복의 효능이 달라지기 때문에 **판매 전략 수립**에 도움
- 기존의 복잡한 나이 예측 방법을 대체할 수 있는 방법론 개발

데이터 설명

01 데이터 크기

- Train : (90615, 10)
- Original : (4177, 10)
- Test : (60411, 9)

02 피처 요약표

	feature	type	Null	unique	sample_0	sample_1	sample_2	dtype
0	Sex	object	0	3	F	F	I	명목형
1	Length	float64	0	157	0.55	0.63	0.16	연속형
2	Diameter	float64	0	126	0.43	0.49	0.11	연속형
3	Height	float64	0	90	0.15	0.145	0.025	연속형
4	Whole weight	float64	0	3175	0.7715	1.13	0.021	연속형
5	Whole weight.1	float64	0	1799	0.3285	0.458	0.0055	연속형
6	Whole weight.2	float64	0	979	0.1465	0.2765	0.003	연속형
7	Shell weight	float64	0	1129	0.24	0.32	0.005	연속형
8	Rings	int64	0	28	11	11	6	연속형

데이터 소개 : 전복의 성별

01 데이터 컬럼

Variable Name	Data Type	Description	Units
Sex	**nominal**	**M, F, and I (infant)**	**-**
Length	continuous	Longest shell measurement	mm
Diameter	continuous	perpendicular to length	mm
Height	continuous	with meat in shell	mm
Whole weight	continuous	whole abalone	grams
Shucked weight (weight.1)	continuous	weight of meat	grams
Viscera weight (weight.2)	continuous	gut weight (after bleeding)	grams
Shell weight	continuous	after being dried	grams
Rings(target)	integer	+1.5 gives the age in years	-

02 설명

- **성별 관련 피쳐**

I : infant

- 대체로 작고 가볍다
- 대체로 어리다

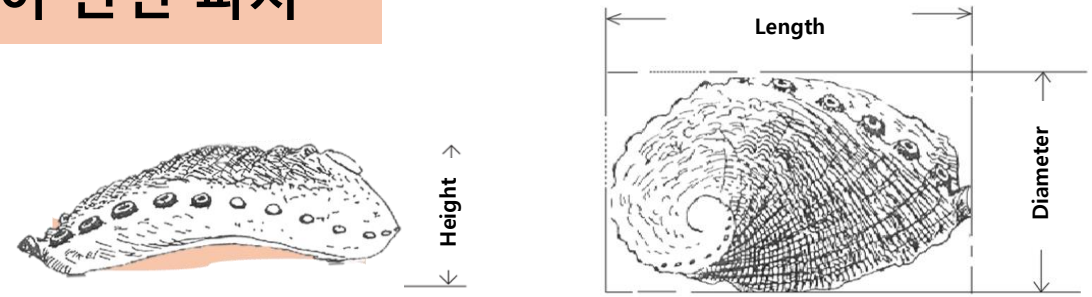
데이터 소개 : 길이/무게 피쳐

01 데이터 컬럼

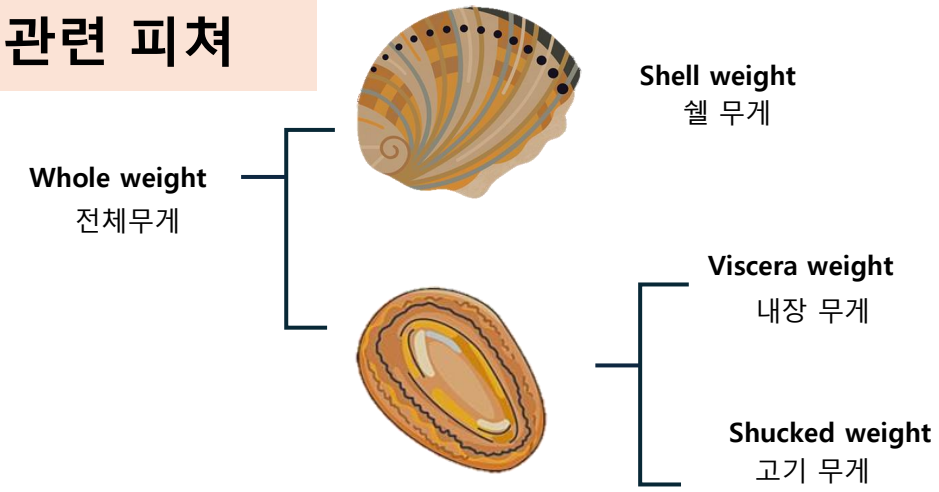
Variable Name	Data Type	Description	Units
Sex	nominal	M, F, and I (infant)	-
Length	**continuous**	**Longest shell measurement**	**mm**
Diameter	**continuous**	**perpendicular to length**	**mm**
Height	**continuous**	**with meat in shell**	**mm**
Whole weight	**continuous**	**whole abalone**	**grams**
Shucked weight (weight.1)	**continuous**	**weight of meat**	**grams**
Viscera weight (weight.2)	**continuous**	**gut weight (after bleeding)**	**grams**
Shell weight	**continuous**	**after being dried**	**grams**
Rings(target)	integer	+1.5 gives the age in years	-

02 설명

- **길이 관련 피쳐**



- **무게 관련 피쳐**



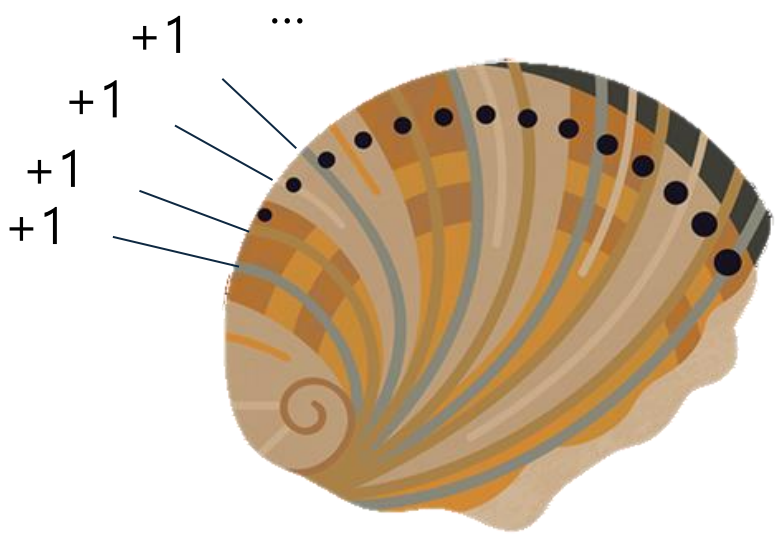
데이터 소개 : 타겟 피쳐

01 데이터 컬럼

Variable Name	Data Type	Description	Units
Sex	nominal	M, F, and I (infant)	-
Length	continuous	Longest shell measurement	mm
Diameter	continuous	perpendicular to length	mm
Height	continuous	with meat in shell	mm
Whole weight	continuous	whole abalone	grams
Shucked weight (weight.1)	continuous	weight of meat	grams
Viscera weight (weight.2)	continuous	gut weight (after bleeding)	grams
Shell weight	continuous	after being dried	grams
Rings(target)	**integer**	**+1.5 gives the age in years**	-

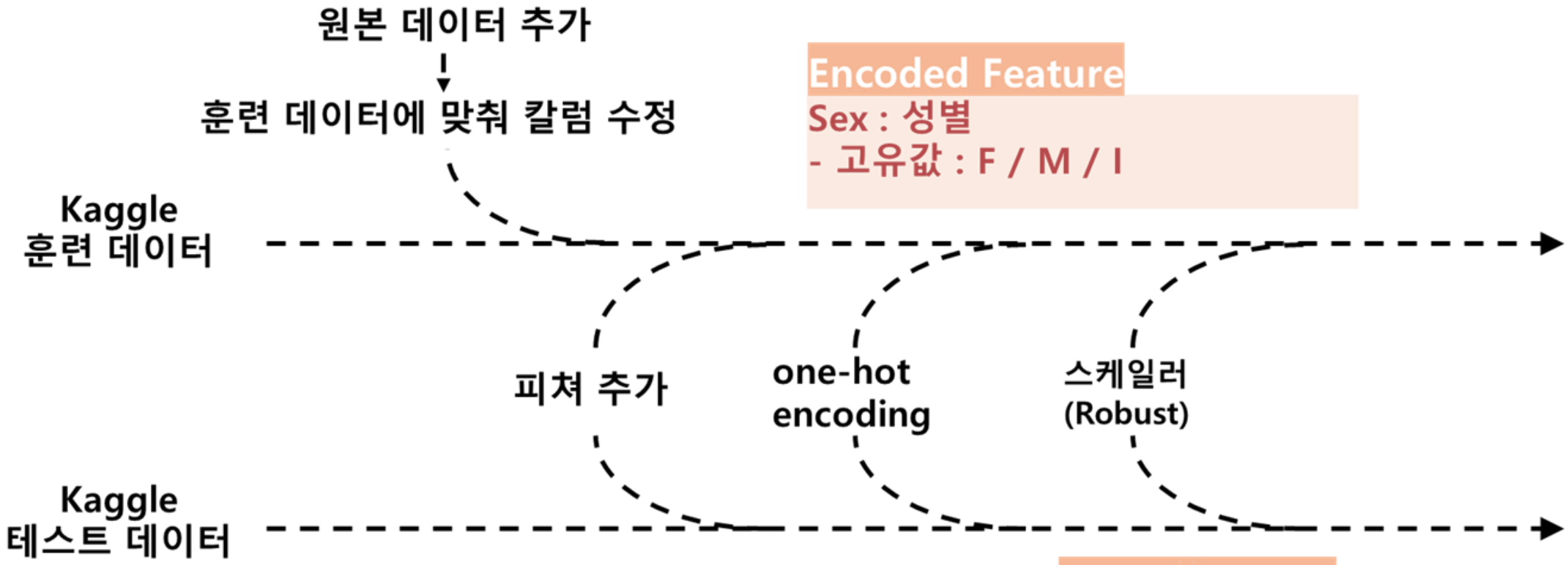
02 설명

- **전복의 나이테 수**



나이테 하나당 1년을 의미,
1.5살을 더해 보정

데이터 전처리



Encoded Feature

Sex : 성별
- 고유값 : F / M / I

추가한 피쳐

- Size : 전복의 크기
- Estimated_grate : 성장속도
- BMI : 전복의 BMI

Scaled Feature

- 길이들 : Length, Diameter, Height
- 무게들 : Whole weight etc.
- 생성 피쳐 : estimated_grate, BMI, size

데이터 EDA

01 데이터 병합

- train data와 original data를 병합해서 EDA 진행

▶ original data

Abalone
Donated on 11/30/1995

Predict the age of abalone from physical measurements

Dataset Characteristics	Subject Area	Associated Tasks
Tabular	Biology	Classification, Regression
Feature Type	**# Instances**	**# Features**
Categorical, Integer, Real	4177	8

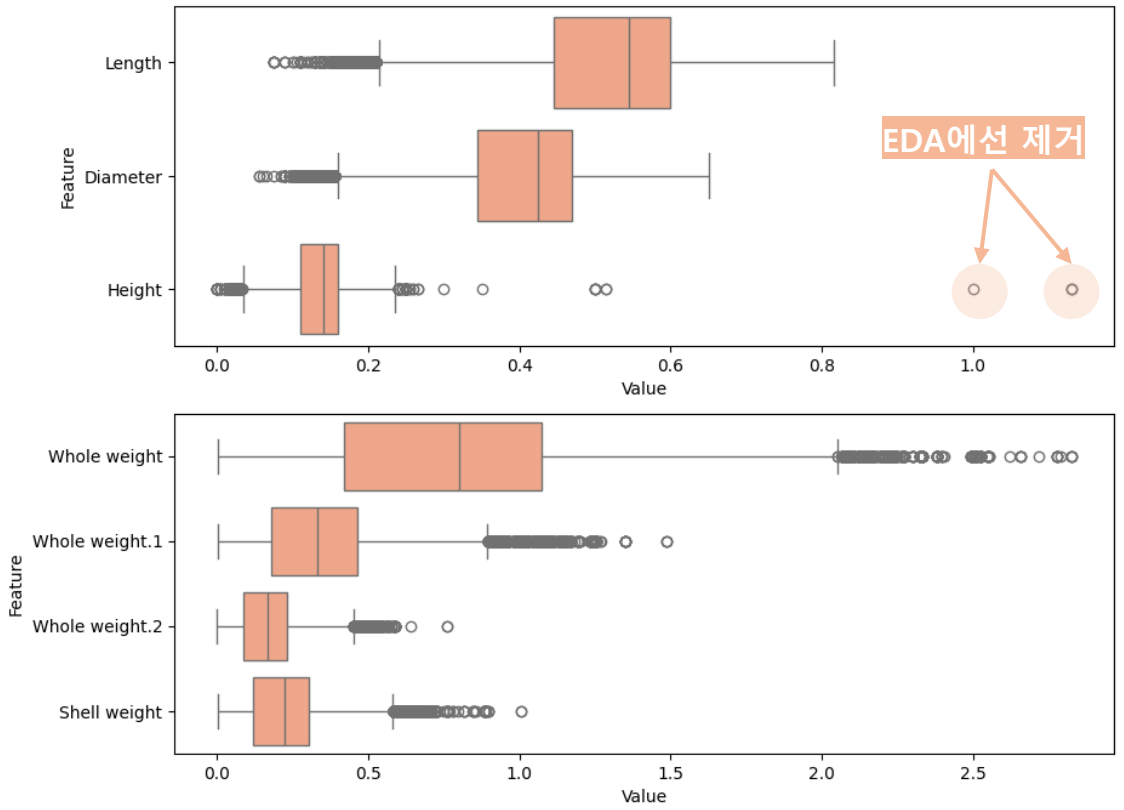
Dataset Information

Additional Information
Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

From the original data examples with missing values were removed (the majority having the predicted value missing), and the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200).

Nash,Warwick, Sellers,Tracy, Talbot,Simon, Cawthorn,Andrew, and Ford,Wes. (1995). Abalone. UCI Machine Learning Repository.

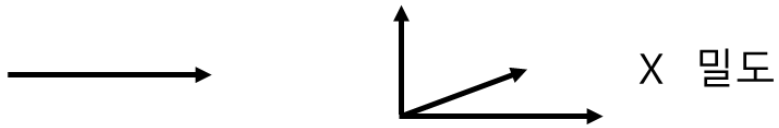
02 데이터 이상치 확인



데이터 상관관계

데이터간 상관관계

- 모든 데이터간 상관관계는 매우 큼



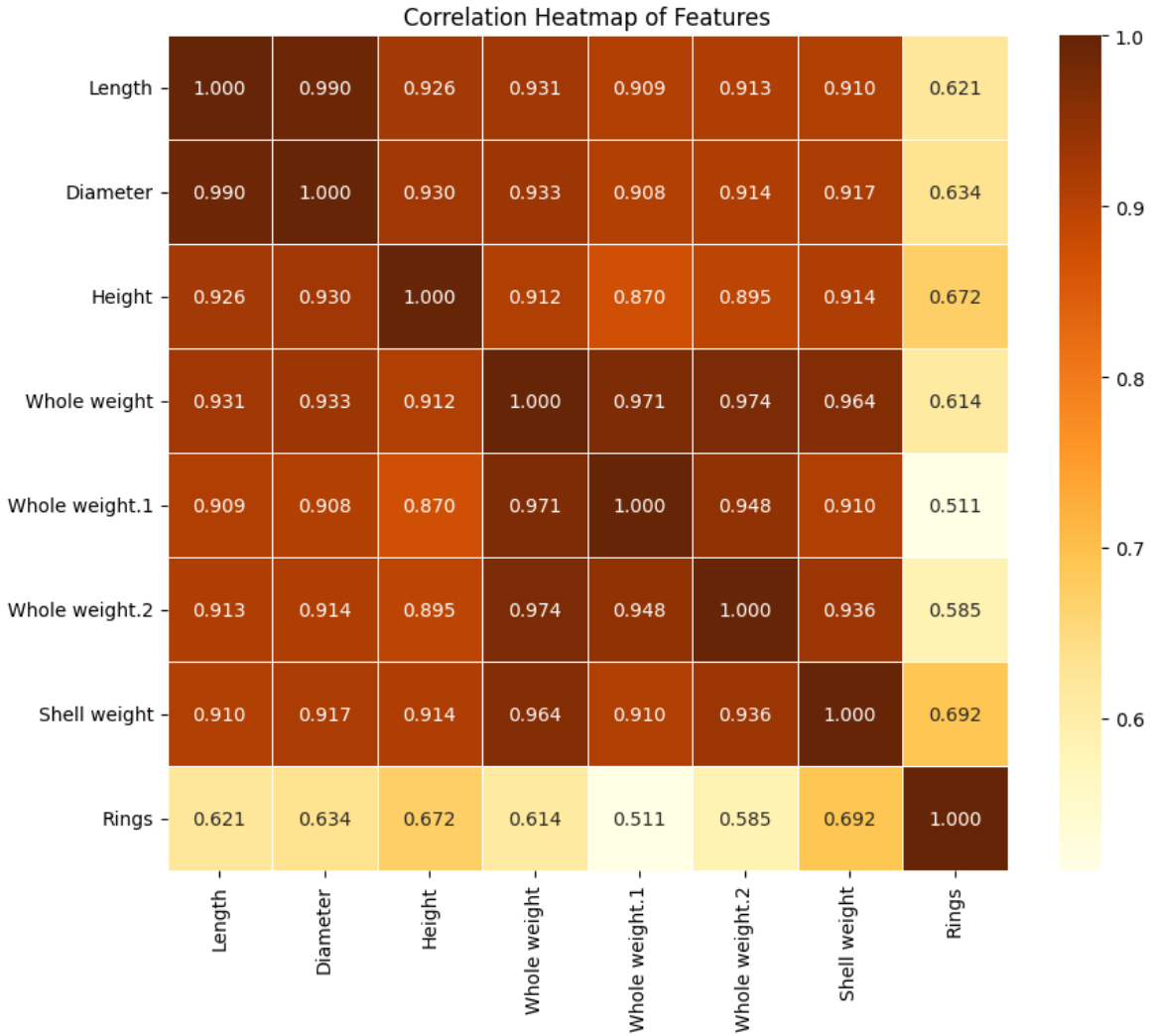
~ $1m$/Ring

길이 증가

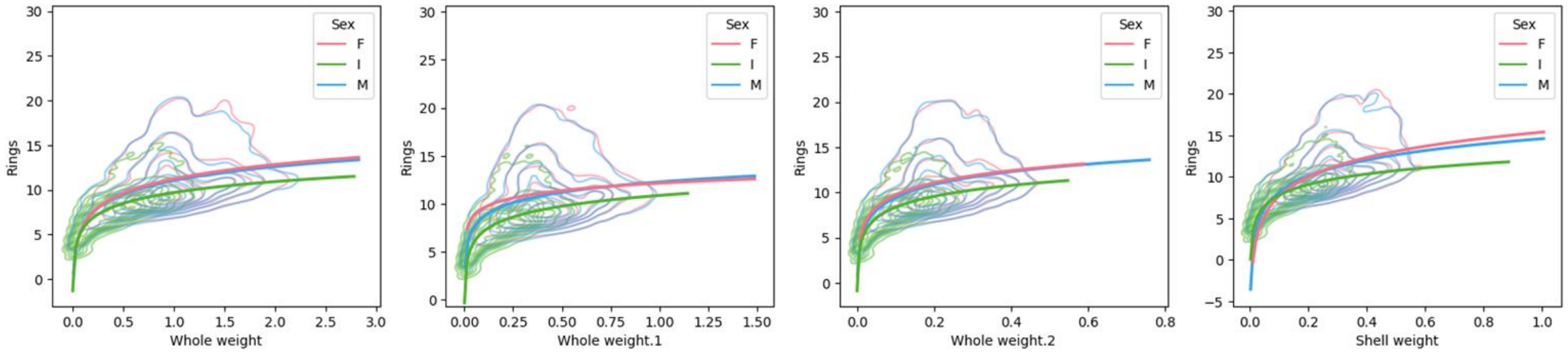
X 밀도

~$1m^3$/Ring

무게 증가

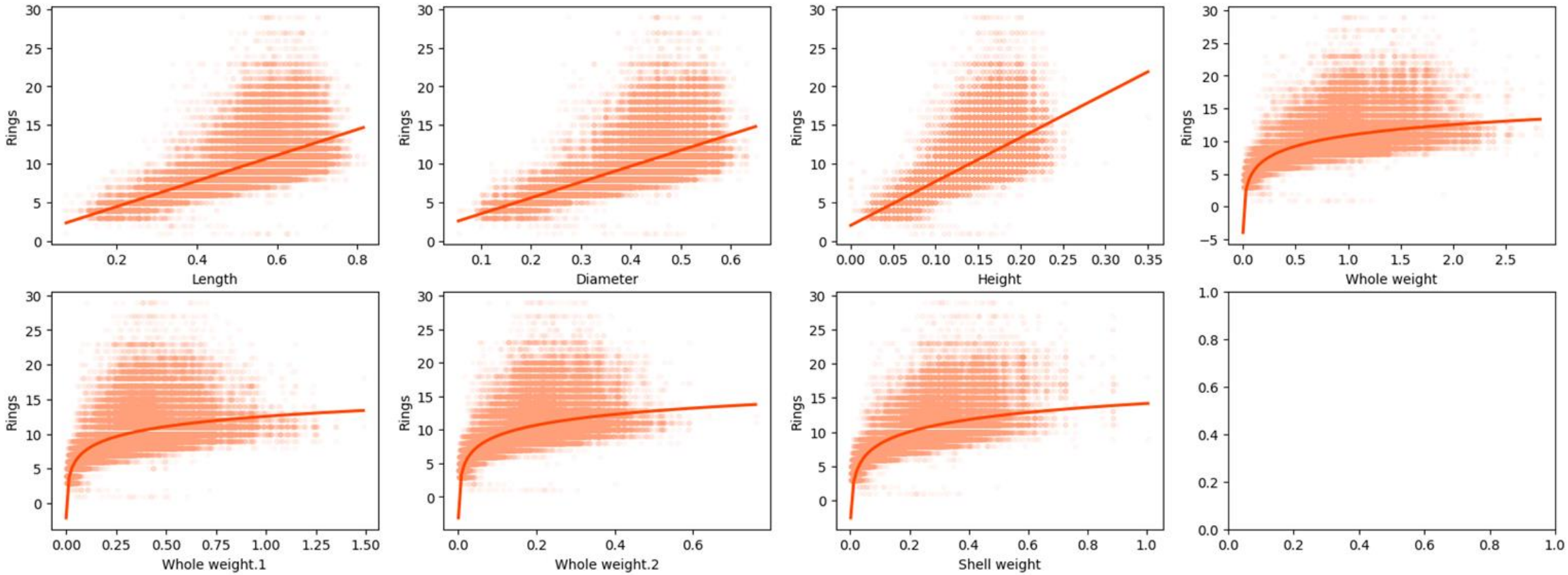


데이터 EDA : 성별에 따른 분포



Female/Male 집단과 Infant 집단 사이에는 회귀상으로도 명확한 차이를 나타냄

데이터 EDA : Regplot

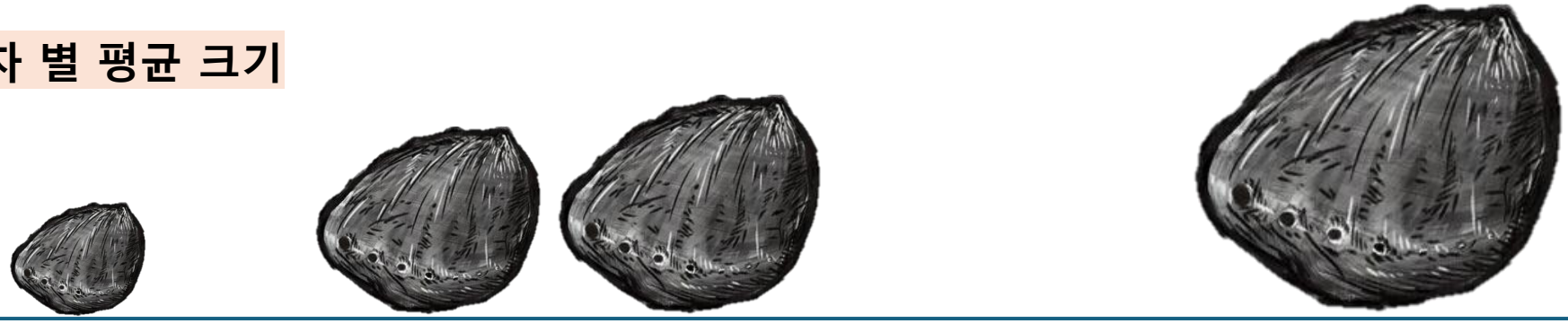


길이 $Rings(\ell[mm]) \sim \ell[mm]$

무게 $Rings(m[g]) \sim \ln(m[g])$

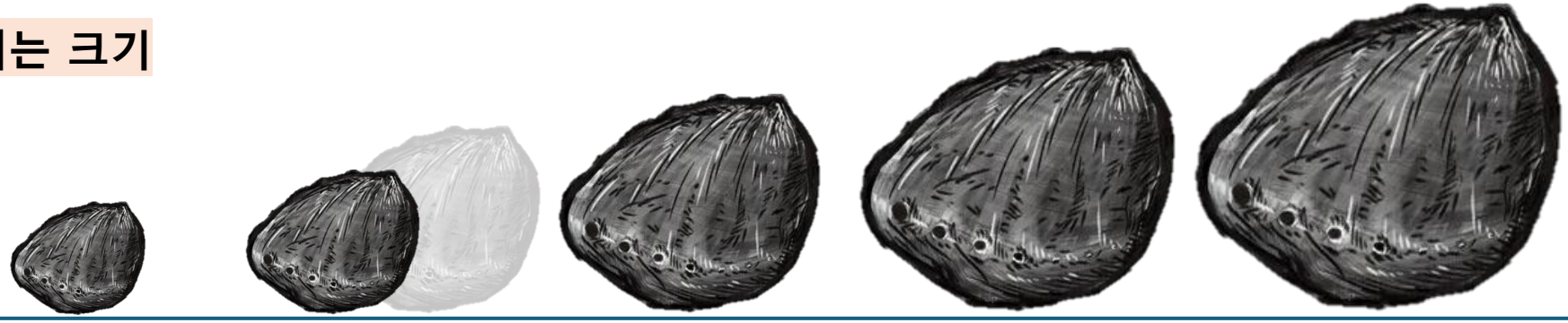
모델 전략 : 성장속도

전복의 년차 별 평균 크기



3 년차 | **6 년차** | 7 년차 | 15 년차

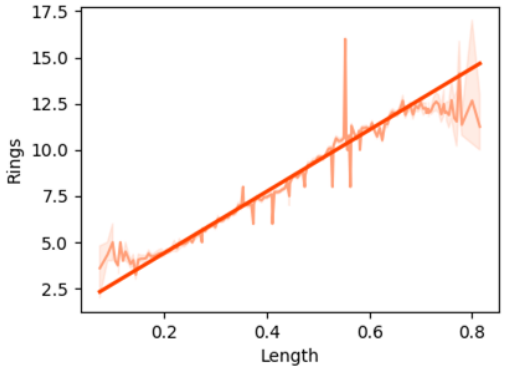
실제 수확되는 크기

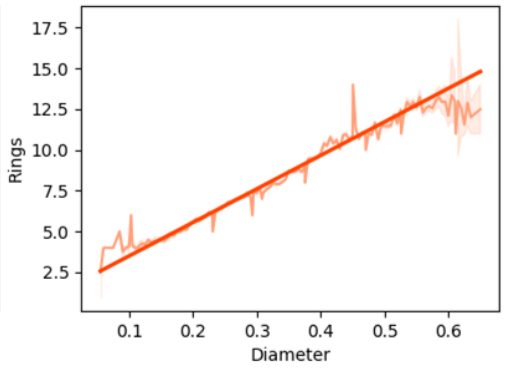


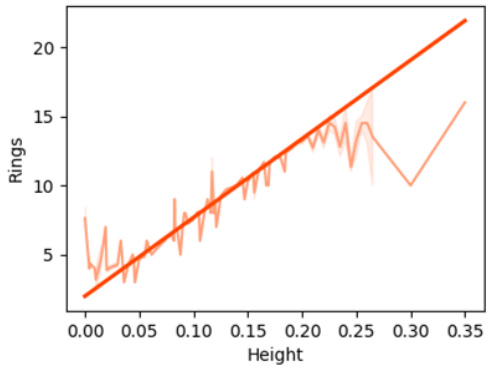
3 년차 평균 | **6 년차** | 7 년차평균 | **6 년차** | 15 년차 평균

각각의 성장 속도가 상이하기때문에 외적인 크기 전복의 나이를 예측할 수 없는 상황

모델 전략 : 성장속도



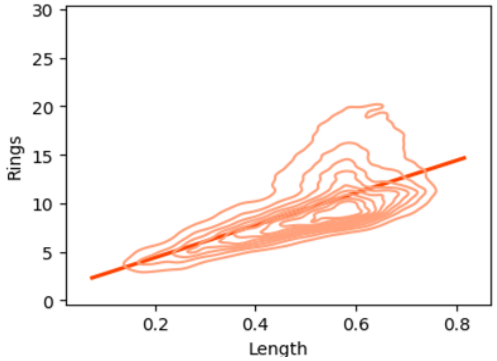


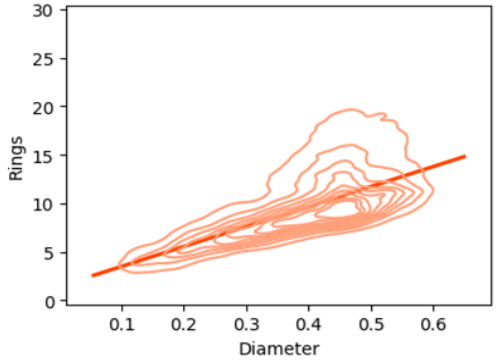


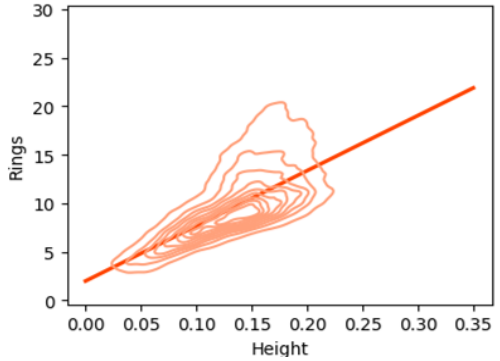
- 평균내면 분명한 상관관계가 확인됨

$$Rings(\ell[mm]) \sim \ell[mm]$$

$$Rings(m[g]) \sim \ln(m[g])$$







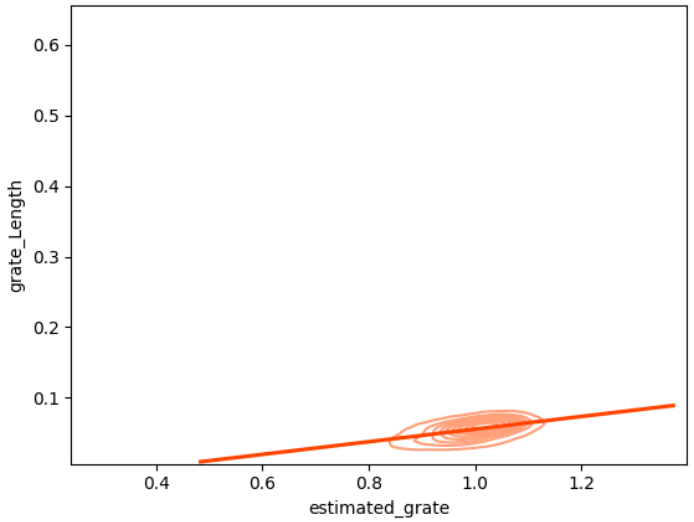
- 성장속도에 의한 변동성이 커서 모델 예측성이 떨어짐

모델 전략 : 성장속도


황교익씨 죽기전에
먹어야할 음식 101
천혜 바다를 머금은 맛! 완도 전복
적어도 2년 반 동안 같이 자란 거죠
TV CHOSUN

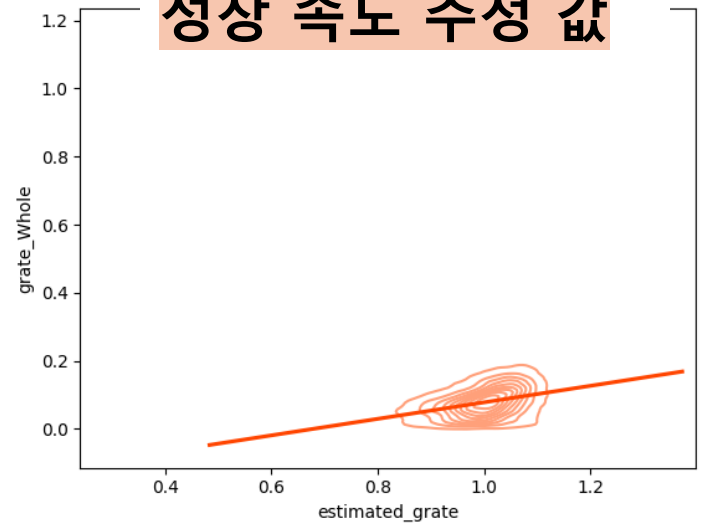
피쳐 생성 : 성장속도 관련 값 | estimated_grate

길이 성장 속도



성장 속도 추정 값

무게 성장 속도



각 전복의 성장 속도는 아래 식과 관련이 있다고 추정

성장 속도 추정 값 ~ (고기 + 내장) / (전체 무게) *

***정확한 식**
(Whole weight.1 + Whole weight.2) / (Whole weight.1 + Whole weight.2 + Shell weight)

피쳐 생성 : BMI

Intestinal microbial diversity is higher in Pacific **abalone** (Haliotis discus hannai) with slower growth rates

MJ Choi, YD Oh, YR Kim, HK Lim, JM Kim - Aquaculture, 2021 - Elsevier

... and large size Pacific **abalone** with higher **BMI** values in this study show they have little impact on **abalone obesity** and growth rate. The greater abundance of Firmicutes in **obese** mice ...

☆ 저장 99 인용 24회 인용 관련 학술자료 전체 5개의 버전

Abalone *Haliotis* spp.

Y Koizumi, Y Tsuji - Application of Recirculating Aquaculture Systems in ..., 2017 - Springer

... In our study, almost the same growth rates (shell length, body weight, and **BMI**) of **abalone** fed with dried Laminaria angustata and artificial diet have been achieved in the same rearing ...

☆ 저장 99 인용 7회 인용 관련 학술자료 전체 3개의 버전

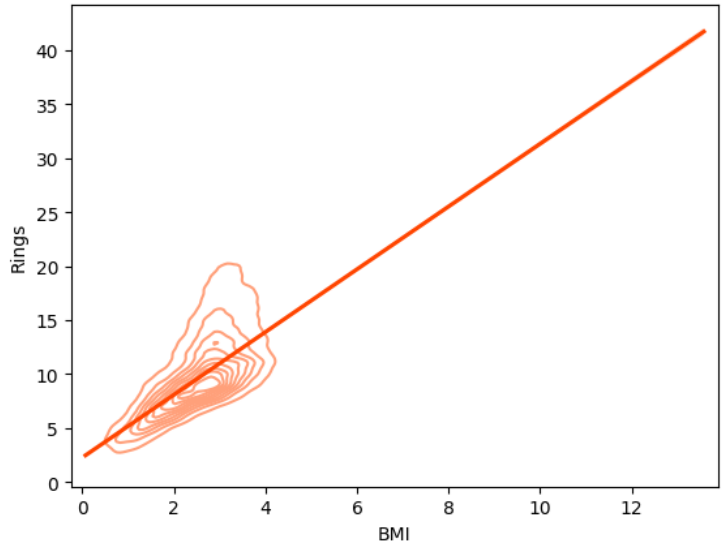
Stock enhancement of **abalone**, Haliotis asinina, in multi-use buffer zone of Sagay Marine Reserve in the Philippines

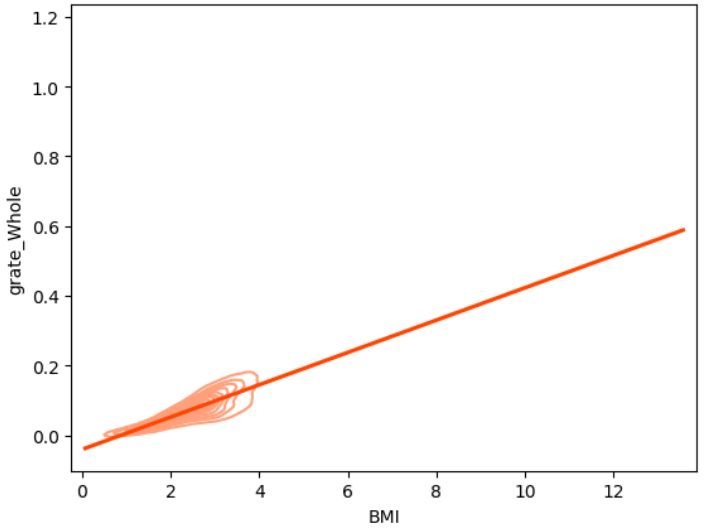
ND Salayo, T Azuma, RJG Castel, RT Barrido... - Aquaculture, 2020 - Elsevier

... mean SL, BW and **BMI** of the HR and wild **abalone** during every sampling period showed ... the t-values of the mean SL, BW and **BMI** of the HR and wild abalones during every sampling ...

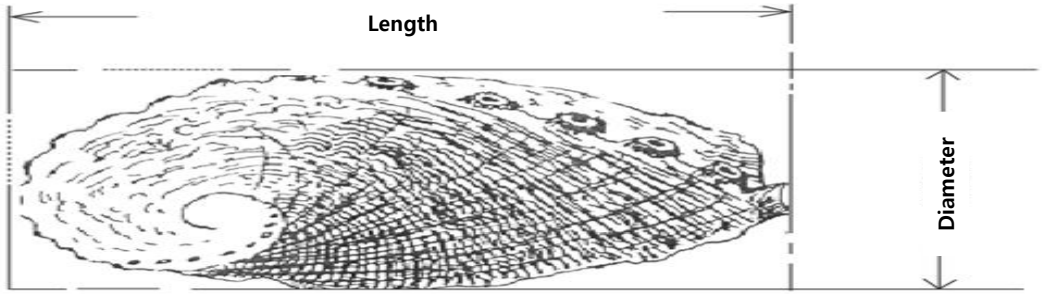
☆ 저장 99 인용 12회 인용 관련 학술자료 전체 8개의 버전

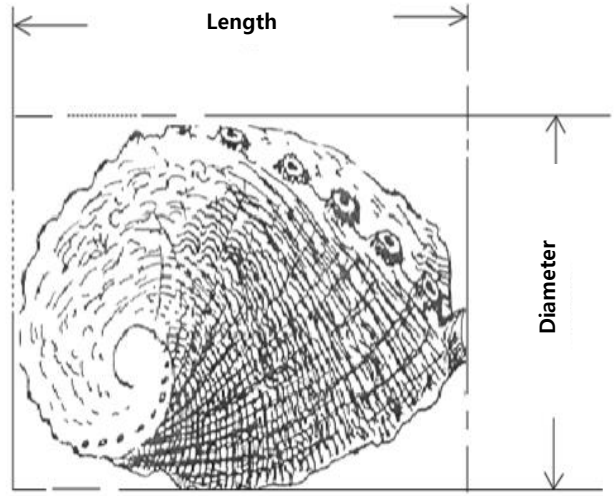
여러 논문에서 전복의 BMI 수치를 빈번하게 구해서 넣어보니 **전복의 나이, 성장 속도와 상관관계**를 보임





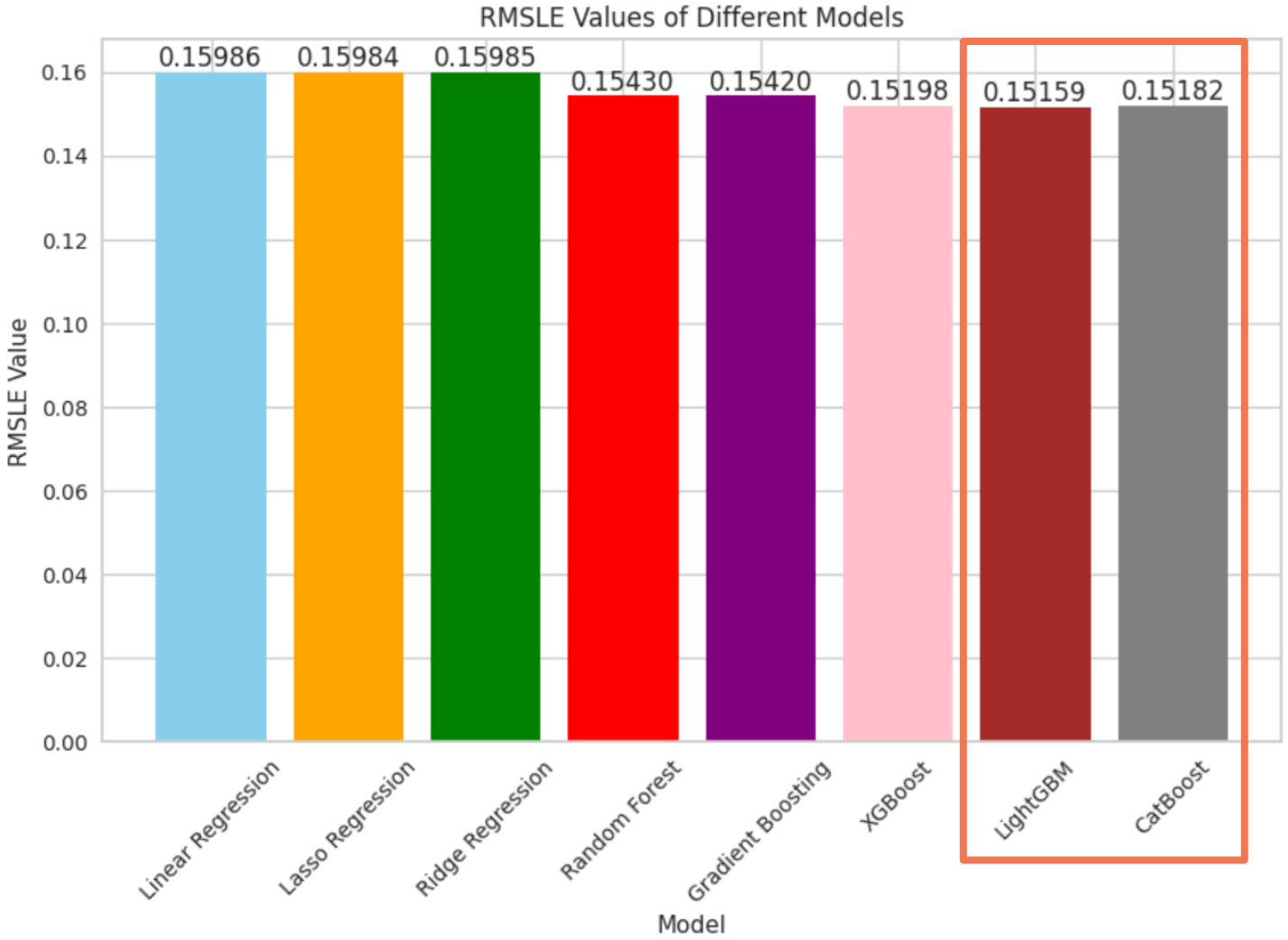
피쳐 생성 : size





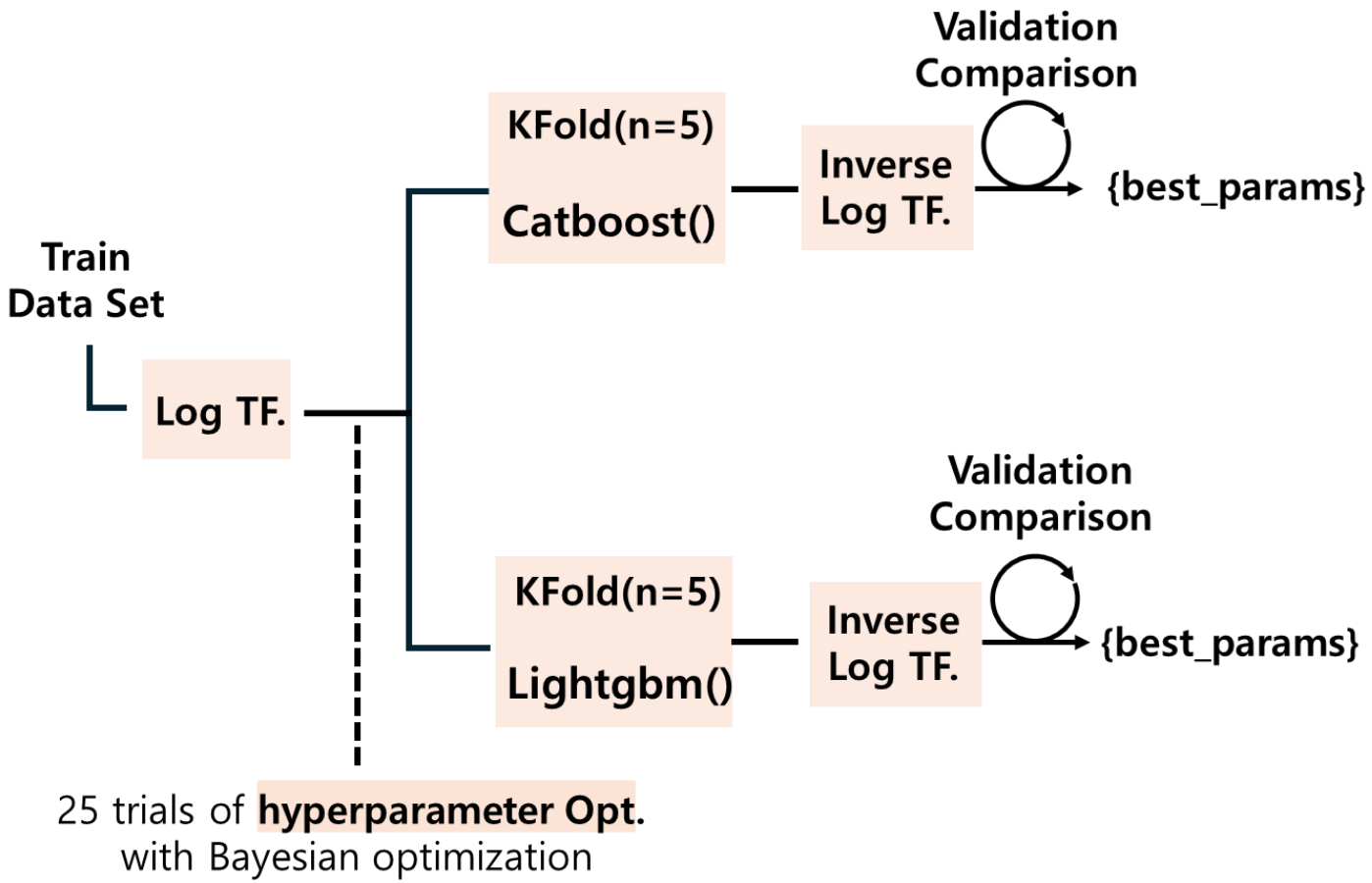
size(= Length X Diameter)가 전복의 크기를 잘 나타냄

baseline 모델 성능 평가



상위 2개 모델로 보팅 진행
Cat, LGBM

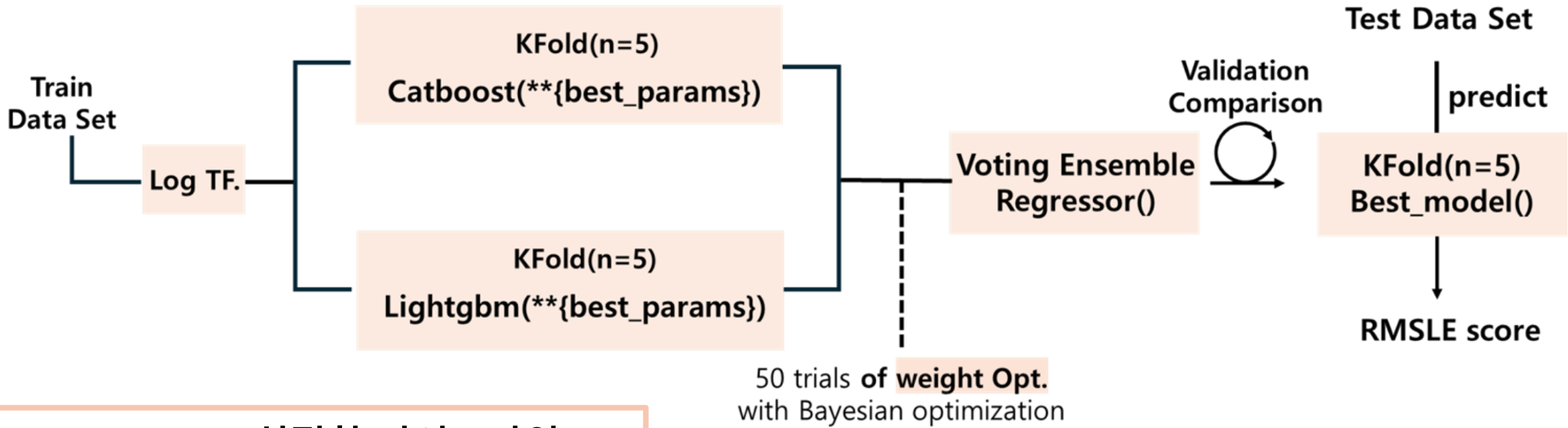
모델 파이프라인 : 하이퍼파라미터 최적화



hyperparameter 최적화 파이프라인

- ✅베이지안 최적화 이용
- ✅KFold 적용
- ✅Log 변환 적용

모델 파이프라인 : voting weight 최적화



voting weight 최적화 파이프라인

- ✅베이지안 최적화 이용
- ✅KFold 적용
- ✅Log 변환 적용

대회 제출 결과

01 대회 순위 결과

▶ 기존 목표 (상위 10% 이내, Score : 0.14564)

(24.04.25 17시 기준)

02 제출 결과

▶ 제출 Score :

▶ 예상 등수 : 등 (24.04.25 17시 기준)

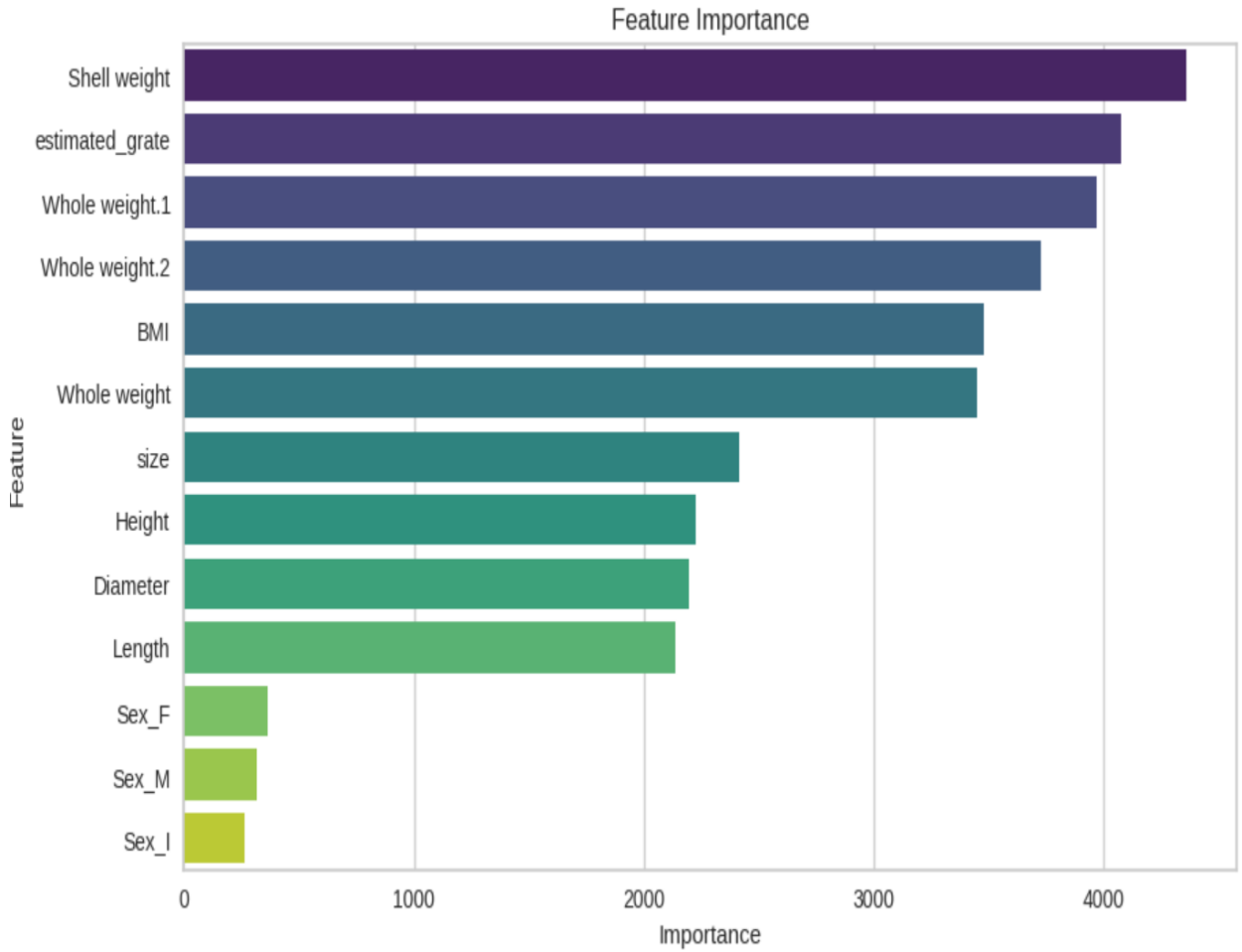
인사이트

01 가설 평가

1) **BMI**는 전복의 나이를 예측함에 있어 의미가 있음

2) **전복의 성장속도(**estimated_grate)는
전복의 나이를 예측함에 있어 높은 중요도를 가짐

3) **전복의 크기(size)**는 전복의 나이를 예측함에 있어
중요도는 낮으나 Length, Diameter보다 중요함



인사이트

02 인사이트 결과

✅전복의 나이는 길이보다 무게와 연관

→ 나이가 많은 전복을 구매하고 싶다면 크기보다 무게에 집중하는 전략

✅전복의 BMI를 고려하면 전복의 나이를 추정하는 데 도움을 줌

→ 전복의 나이에 따라 영양성분이 달라지므로 효능에 맞게 분류&판매 가능

감사합니다.

Q&A