Data Mining

Abalone Data

| 2017110505 김나형 |
| 2017110526 유혜림 |
| 2017110530 홍지연 |

INDEX

1 분석 배경 및 목적

2 데이터 전처리 과정

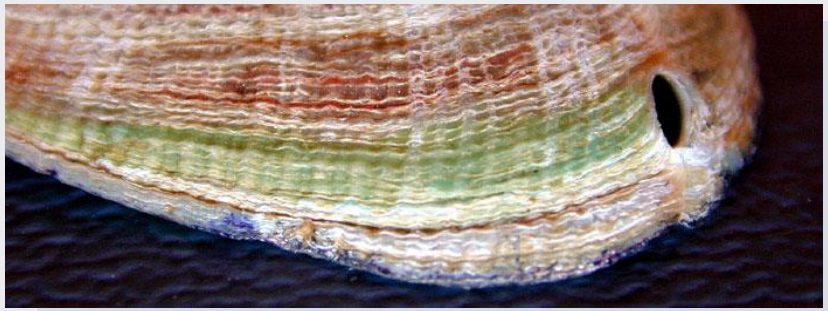
1. 변수 단위 회귀
2. 비논리적 관측치 대체 및 제거
3. 변수 생성
4. 파생 변수 이상치 제거
5. 타깃 변수 범주화
6. 변수 선택 및 데이터 변환

3 모델링

4 결론

분석 배경 및 목적

1 분석 배경 및 목적



전복의 나이는 껍질의 윤문 수로 측정

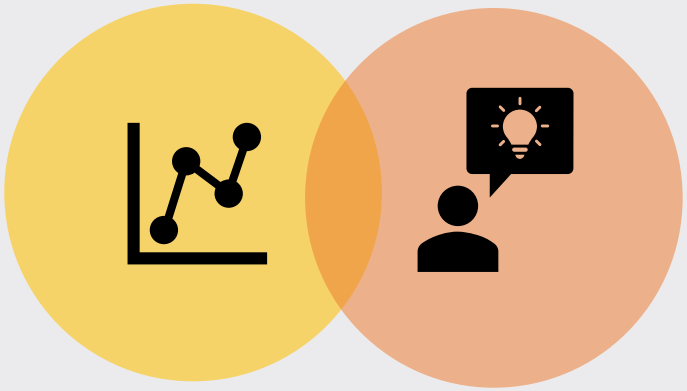
But 과정이 복잡

대신, 길이와 무게로

전복의 성장 정도 추론 가능

1 분석 배경 및 목적

전복 양식사업이 활발한 Tasmania 지역 1차 산업 수산부의 데이터를 이용



전복 데이터 전복 양식업자

1. 분석목적

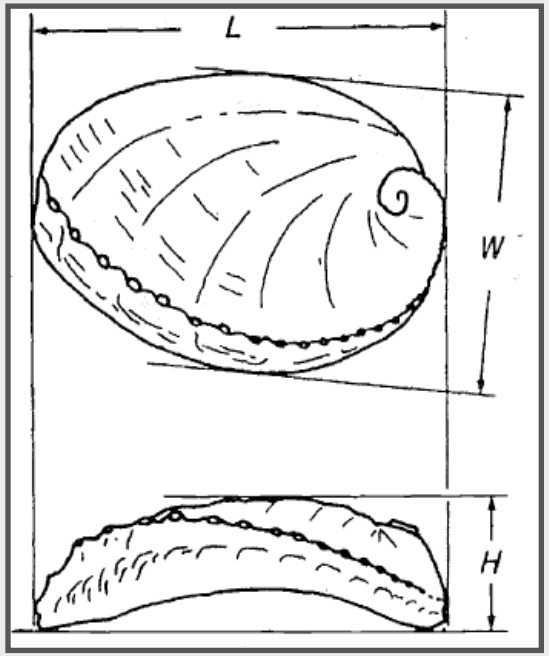
1. 전복의 성장 정도 예측
2. 전복 양식사업의 이익창출을 위한 인사이트 도출

1 분석 배경 및 목적

2. 데이터 소개

구성 : 4177행x9개 변수

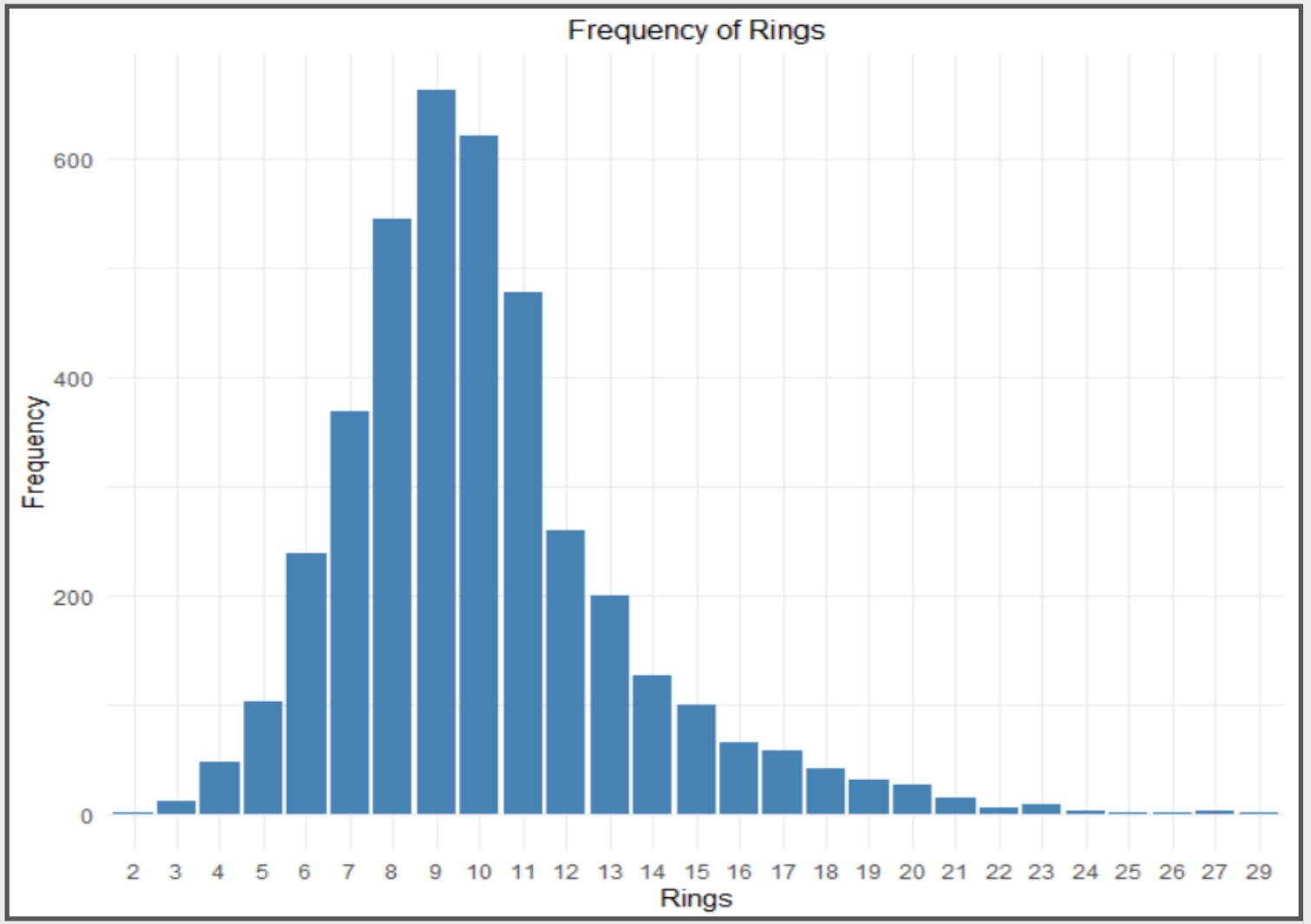
변수명	변수설명	단위
Sex	전복의 성별	
Length	전복 껍질에서 가장 긴 부분의 길이	mm
Diameter	전복 껍질에서 Length의 수직 방향의 길이	mm
Height	전복 껍질과 속살을 포함한 전복의 두께	mm
Whole weight	전복 전체의 무게	g
Shucked weight	내장을 제외한 전복 살의 무게	g
Viscera weight	피를 제거한 내장의 무게	g
Shell weight	물기를 제거한 껍질의 무게	g
Rings	전복의 껍질에 나타난 윤문 수	



1 분석 배경 및 목적

2. 데이터 소개

윤문수에 따른
개체수의 불균형이 심함




2
데이터 전처리

2 데이터 전처리

1. 변수 단위 회귀

Length	Diameter	…	Viscera weight	Shell weight
0.455	0.365	…	0.1010	0.150
0.350	0.265	…	0.0485	0.070
…	…	…	…	…
0.625	0.485	…	0.2610	0.2960
0.710	0.555	…	0.3765	0.4950

X 200

Length	Diameter	…	Viscera weight	Shell weight
91	73	…	20.2	30
70	53	…	9.7	14
…	…	…	…	…
125	97	…	52.2	59.2
142	111	…	75.3	99.0

2 데이터 전처리

2. 비논리적 관측치 제거

- 전복의 두께 (Height)가 0인 경우 (2개)

OBS	Sex	Height	Whole weight	...	Shell weight	Rings
1258	I	0	25.3	...	23.0	8
3997	I	0	26.8	...	70.1	6

OBS	Sex	Height	Whole weight	...	Shell weight	Rings
1258	I	23.208	25.3	...	23.0	8

1258행 같은 성별, 같은 Rings 내 평균으로 대체

3997행 삭제

전복의 전체 무게 < 전복의 껍질 무게
논리적 오류

2 데이터 전처리

2. 비논리적 관측치 제거

- 전복의 전체 무게 < 전복 살 무게 + 전복 내장 무게 + 전복 껍질 무게 인 경우 (159개)

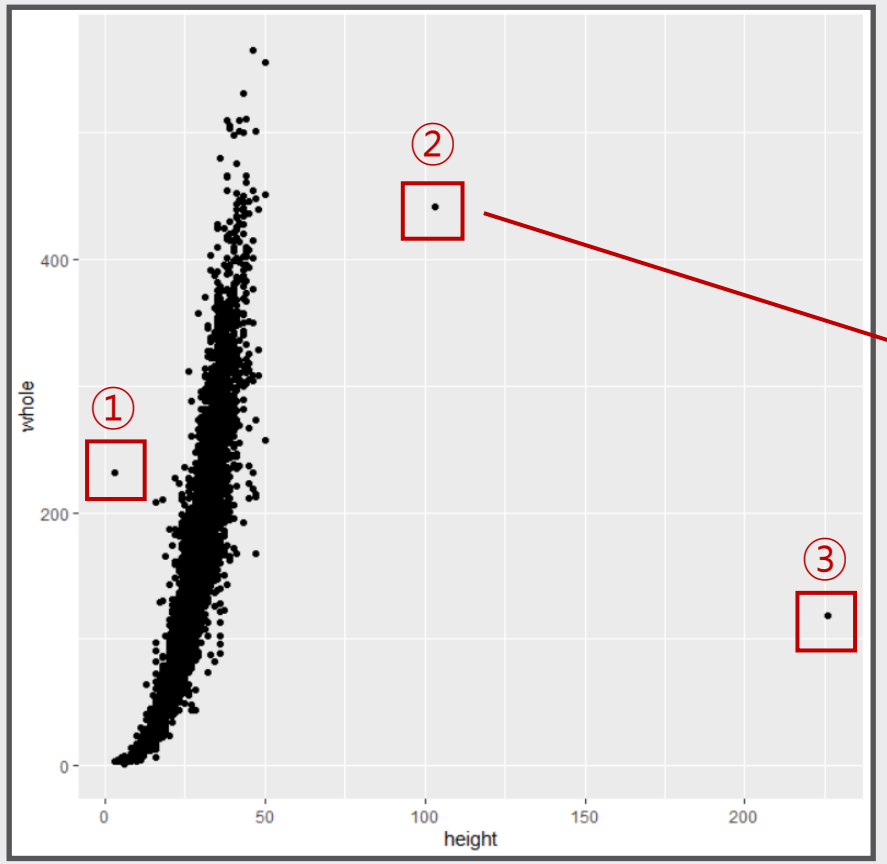
OBS	Sex	…	Whole weight	Shucked Weight	Viscera weight	Shell weight	살 + 내장 +껍질 무게의 합
43	I	…	14.0	6.3	4.7	4.0	15
44	I	…	8.4	5.1	3.0	2.4	10.5
…	…	…	…	…	…	…	…
4047	M	…	133.1	57.0	29.8	53.8	140.6
4144	F	…	271.8	128.4	65.1	81.0	274.5

159개 행 삭제

논리적 오류

2 데이터 전처리

2. 비논리적 관측치 제거 – 그래프상 이상치 (3개)



OBS	Sex	Length	…	Height	Whole weight	Shell weight
1175	F	127	…	3	231.3	102.3
1418	M	141	…	103	442	221.5
2052	F	91	…	226	118.8	66.4

① 1175행: 두께(height)만 극히 작은 값을 보임
② 1418행: 원래 큰 전복으로 보이지만 무게(weight)가 특히 더 큰 값을 보임
③ 2052행: 두께(height)만 극히 큰 값을 보임

2 데이터 전처리

2. 비논리적 관측치 제거 – 그래프상 이상치 (3개)

① 1175행: 성별이 정해진 전복 (Sex가 M,F임) 중 같은 rings내 제 3사분위수 값으로 대체

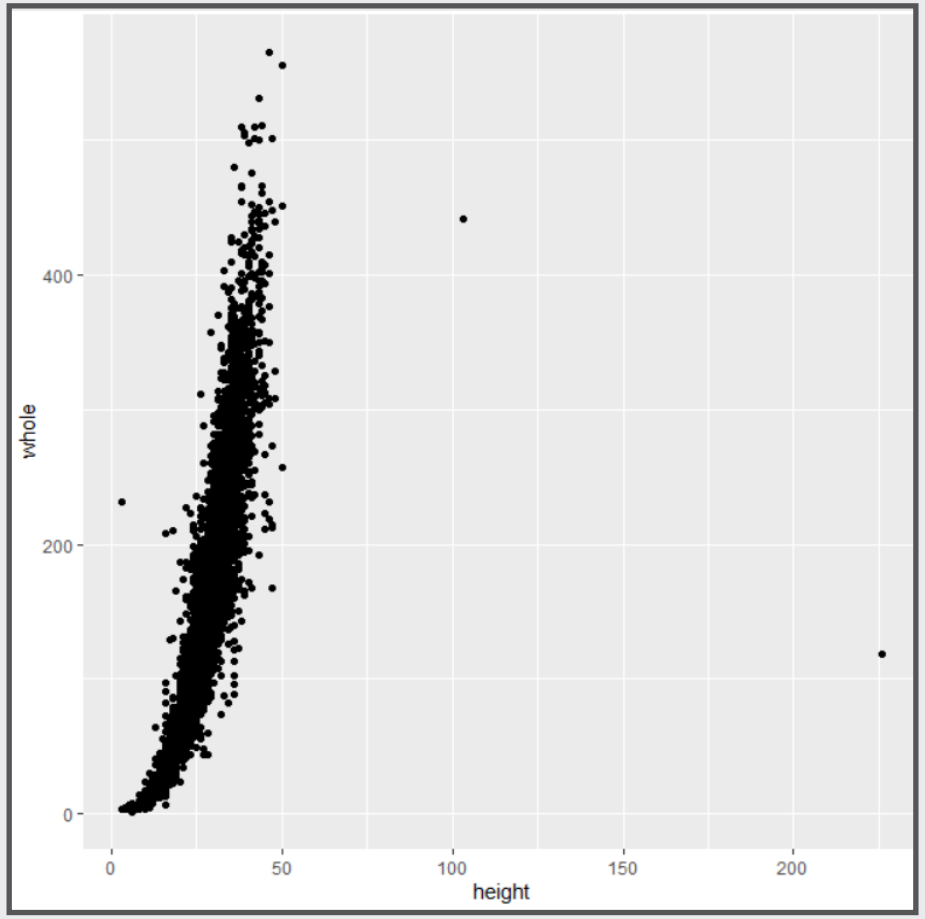
②1418행: 제거

③ 2052행: 성별이 정해진 전복 (Sex가 M,F임) 중 같은 rings내 제 1사분위수 값으로 대체

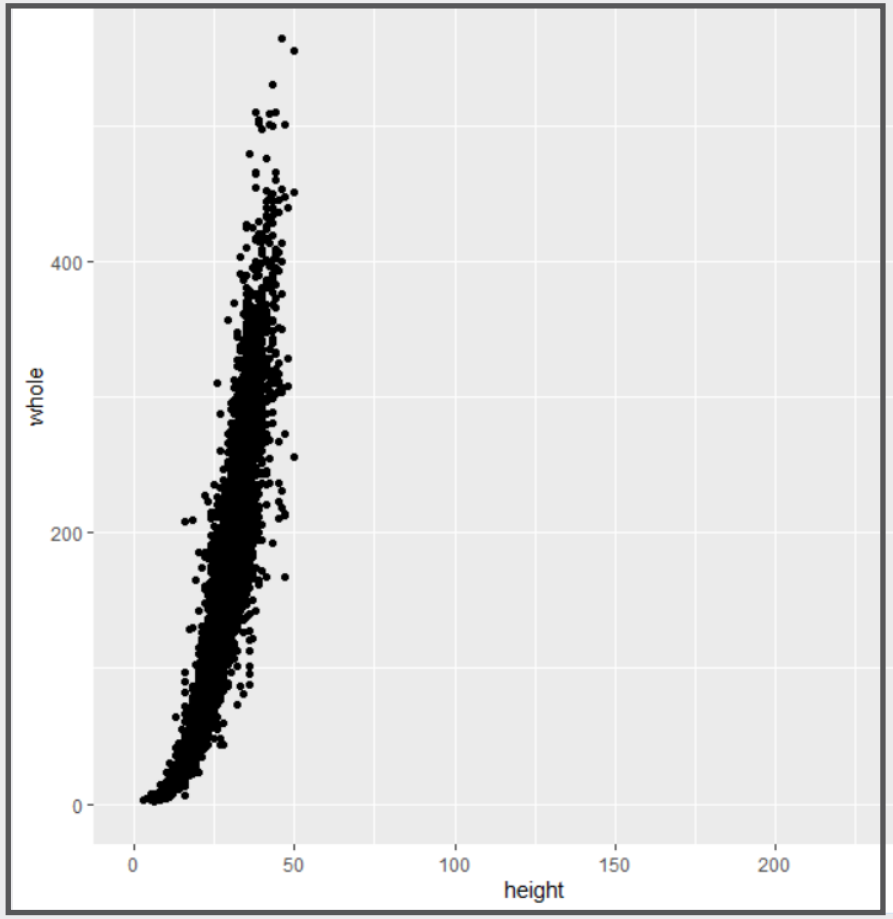
OBS	Sex	Length	…	Height	Whole weight	Shell weight
1175	F	127	…	33	231.3	102.3
2052	F	91	…	24	118.8	66.4

2 데이터 전처리

2. 비논리적 관측치 제거 – 그래프상 이상치 (3개)

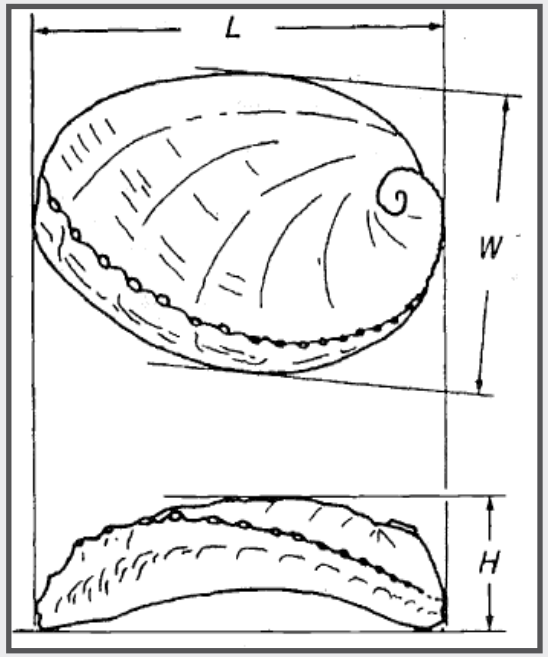


이상치 대체 확인



2. 비논리적 관측치 제거

- Length < Diameter 인 경우 (1개)
- Length : 전복의 껍질에서 가장 긴 길이



OBS	Sex	Length	Diameter	…
43	I	37	75	

OBS	Sex	Length	Diameter	…
43	I	75	37	

두 변수 값 교체

2 데이터 전처리

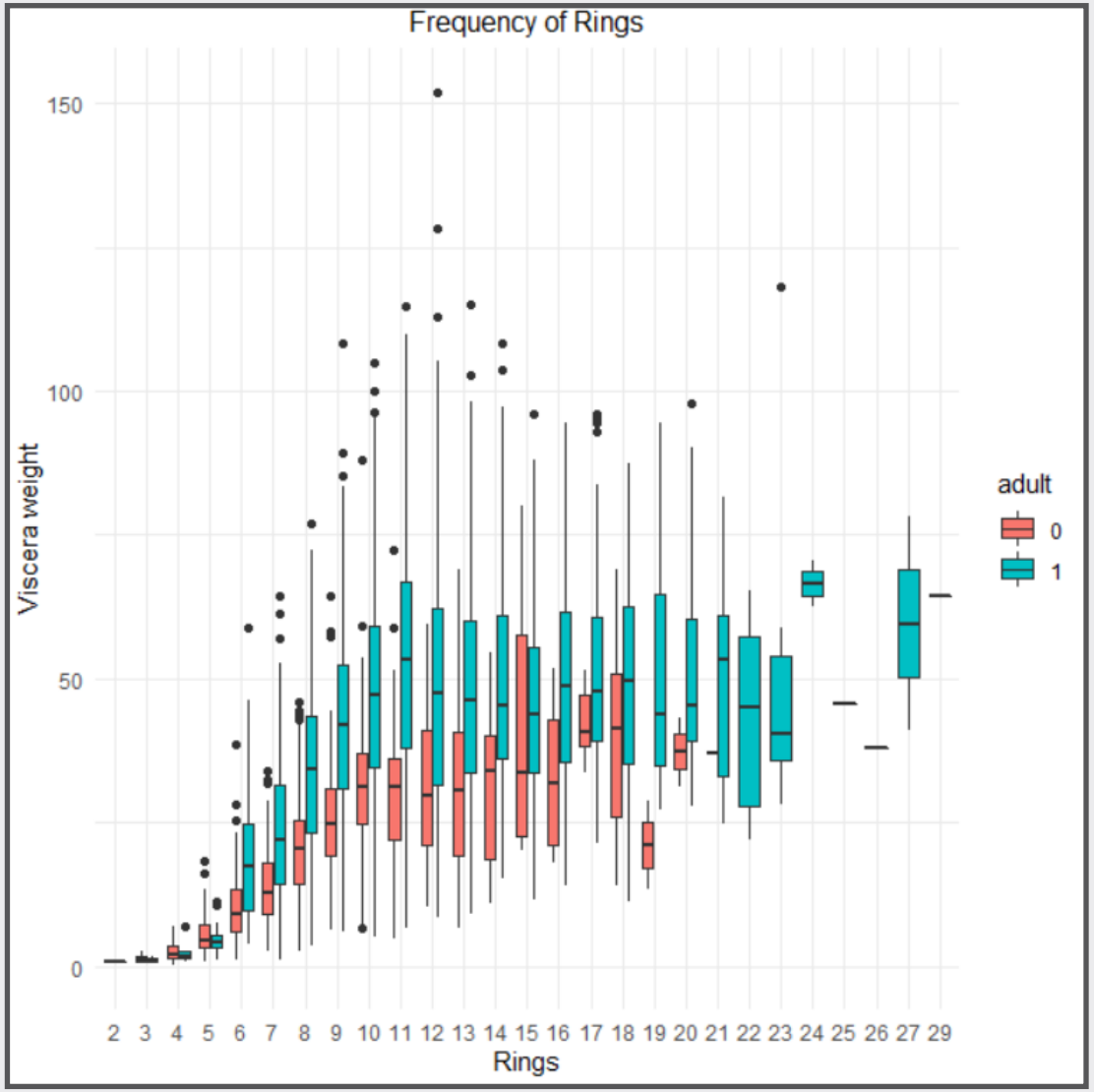
3. 변수 생성 - Adult

Infant

- 실제로 어린 전복
- 나이가 많거나 크기가 큼에도 내장이 제대로 형성되지 않아 성별을 논할 수 없는 전복

Female, Male ➔ Adult=1
Infant ➔ Adult=0

➔ 성별 결정 여부에 따라 내장의 무게를 비교해본 결과 치패의 내장 무게는 성별이 결정된 전복의 무게보다 덜 나가는 것을 확인



2 데이터 전처리

3. 변수 생성

변수명	변수 설명	계산식
Adult	전복의 성숙 (성인 여부)	Female, Male ➔ 1 / Infant ➔ 0
Volume	전복의 부피	$\frac{4}{3} \times \pi \times \frac{\text{length}}{2} \times \frac{\text{diameter}}{2} \times \frac{\text{height}}{2}$
Shell size	껍질의 크기 (껍질의 가장 긴 길이를 포함하는 단면)	$\pi \times \frac{\text{length}}{2} \times \frac{\text{diameter}}{2}$
Whole Unit	단위 부피 당 전체 무게	$\frac{\text{Whole weight}}{\text{Volume}}$
Shucked unit	단위 부피 당 고기(살)의 무게	$\frac{\text{Shucked weight}}{\text{Volume}}$

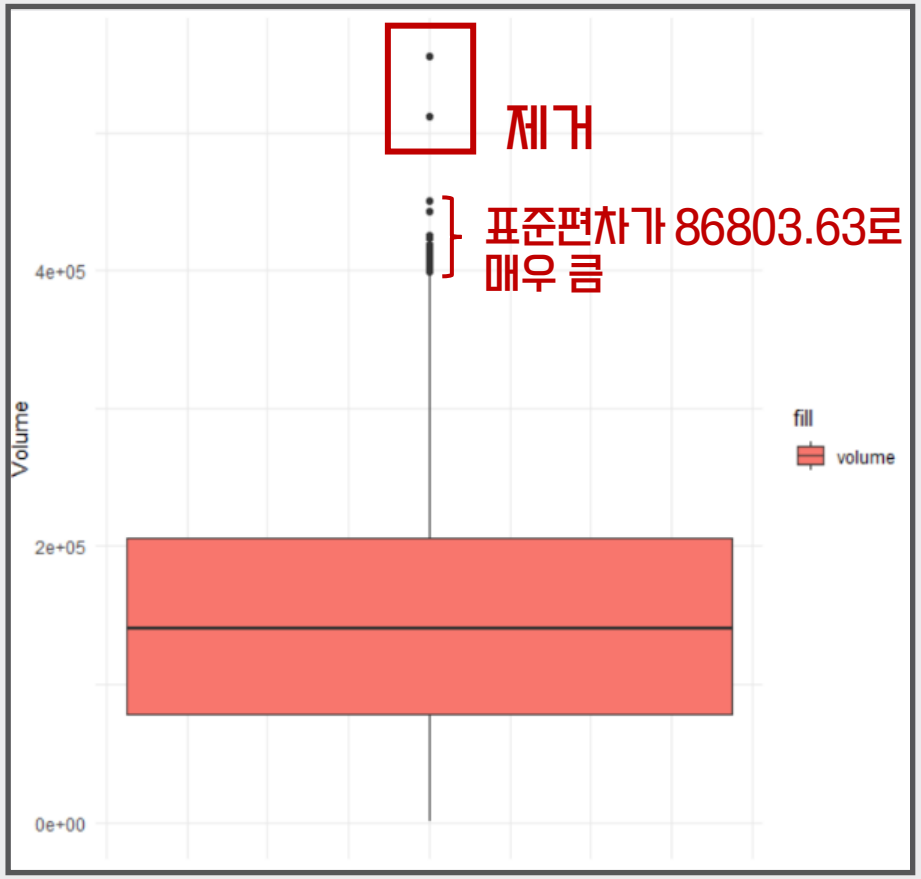
2 데이터 전처리

3. 변수 생성

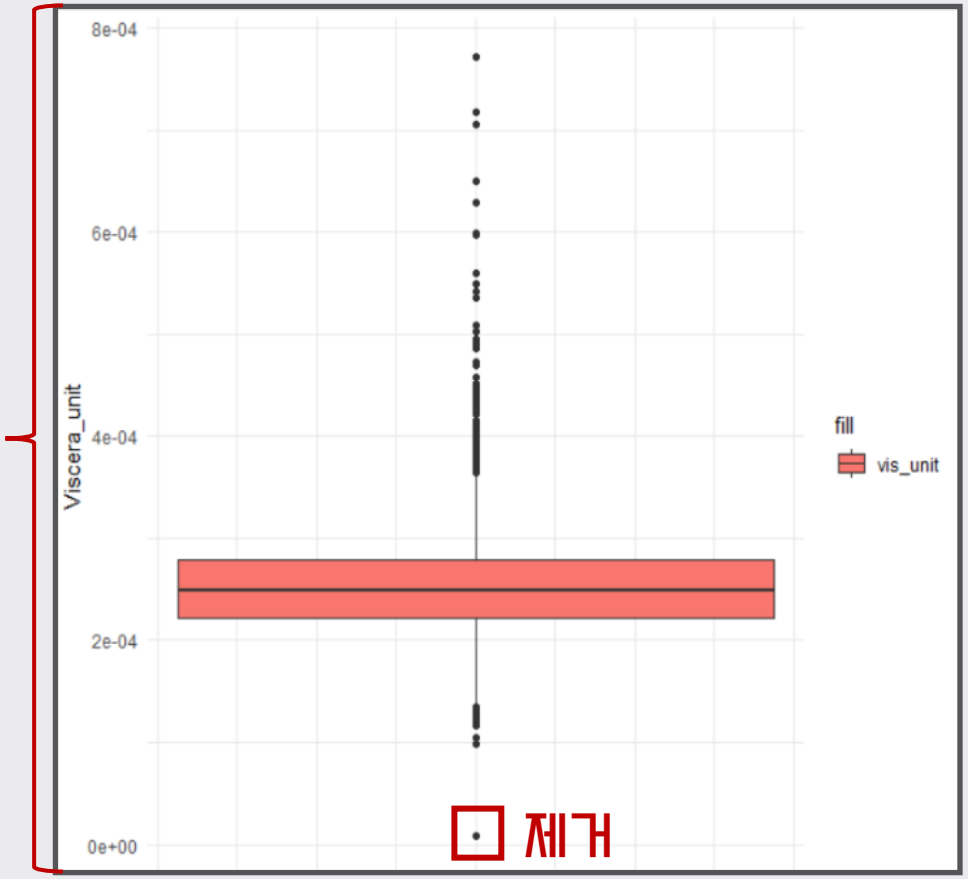
변수명	변수 설명	계산식
Viscera Unit	단위 부피 당 내장의 무게	$\frac{\text{Viscera weight}}{\text{Volume}}$
Shell ratio	단위 면적 당 껍질의 무게	$\frac{\text{Shell weight}}{\text{Shell size}}$
Weight ratio	전체 무게에서 내장과 살이 차지하는 비율	$\frac{\text{Shucked weight+Viscera weight}}{\text{Volume}}$
Size growth	연 평균 껍질의 성장 정도	$\frac{\text{Shell size}}{(\text{rings}+1.5)}$

2 데이터 전처리

4. 파생 변수 이상치 제거 – 전복의 부피 (Volume), 단위 부피 당 내장 무게 (Viscera_unit)

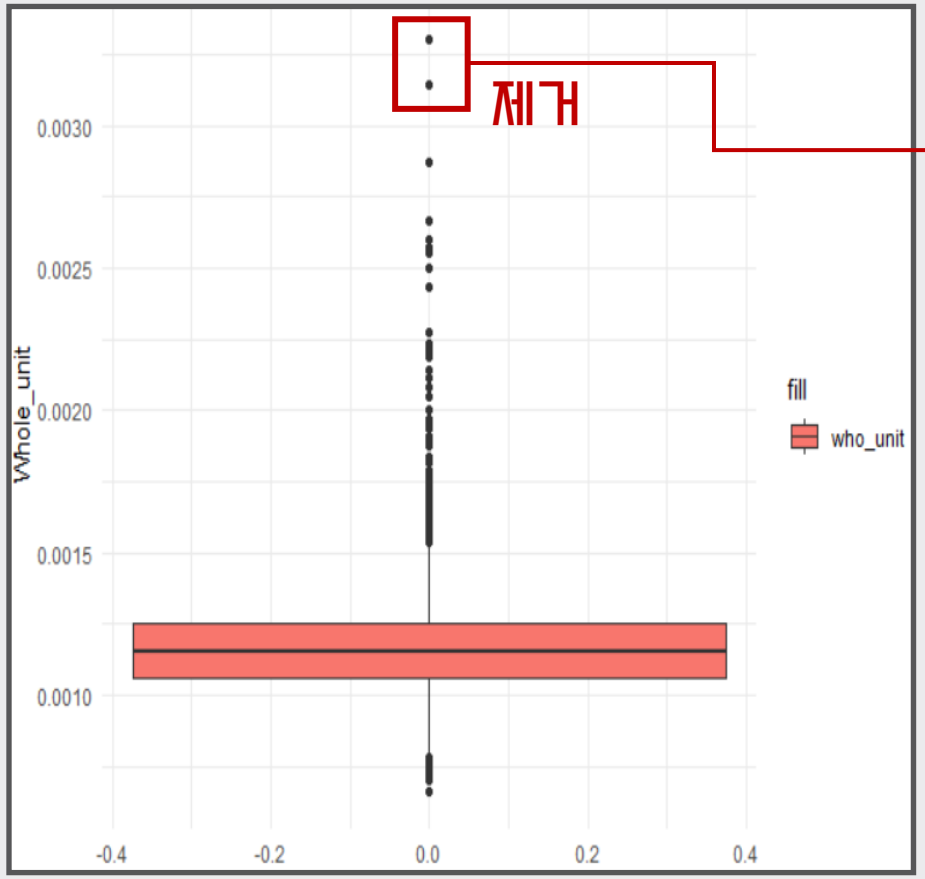


값이 매우 작음

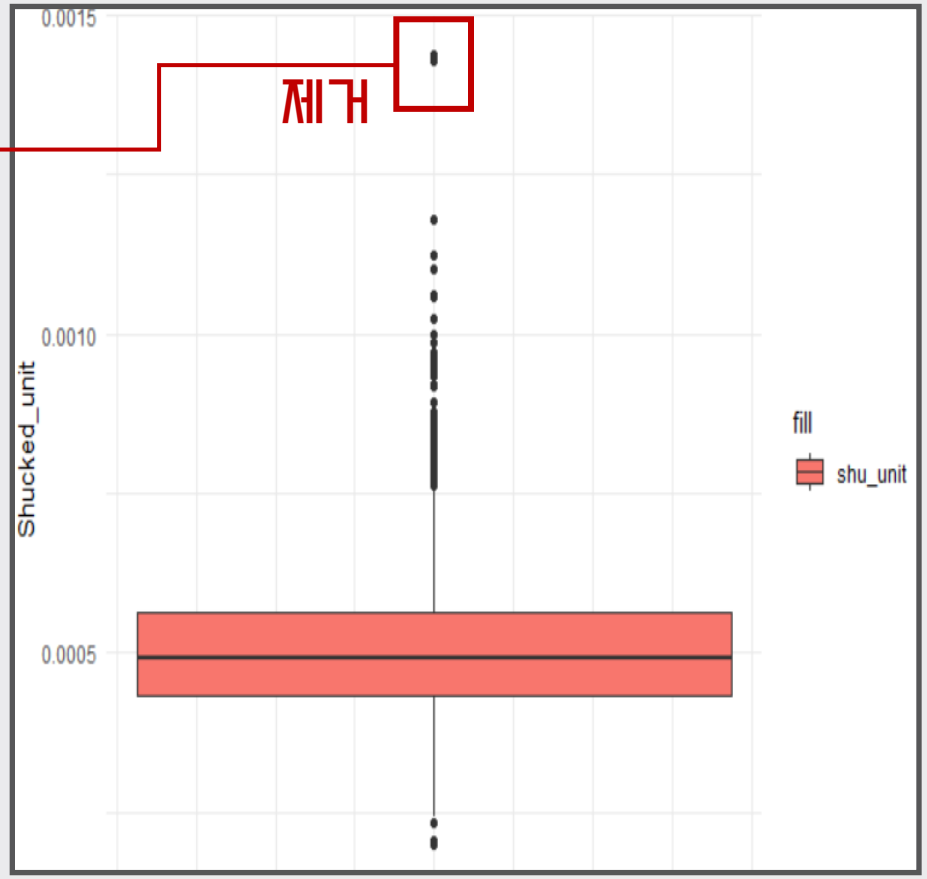


2 데이터 전처리

4. 파생 변수 이상치 제거 - 단위 부피 당 전체 무게와 살의 무게 (Whole_unit , Shucked_unit)



두 점이 각각
같은 관측치

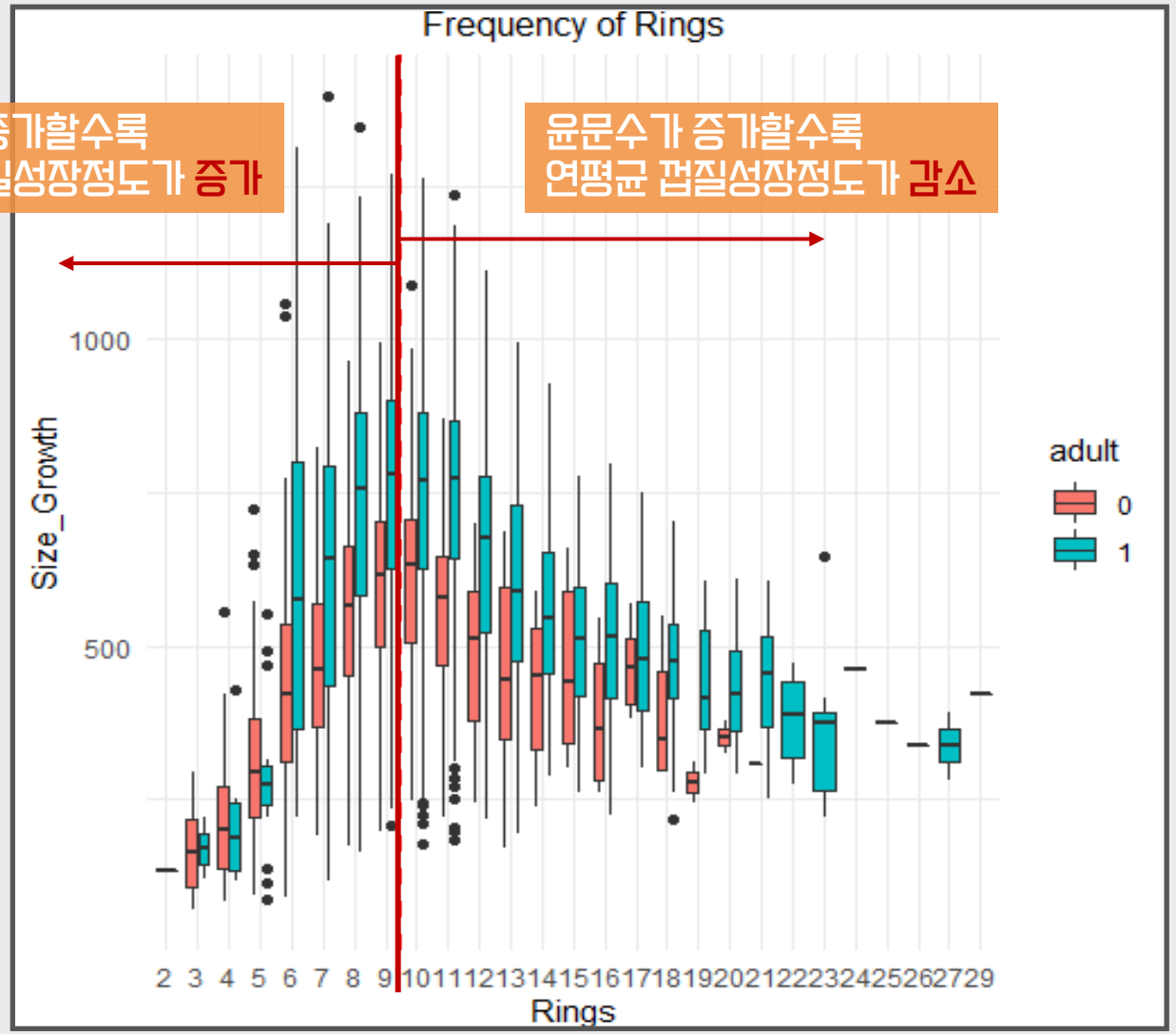


5. 타깃 변수 범주화

Rings 2~9 ➔ Status = 0
Rings 10~29 ➔ Status = 1

Status가 0인 전복은 성장 가능성이 ↑
➔ 양식을 지속할 가치가 있음

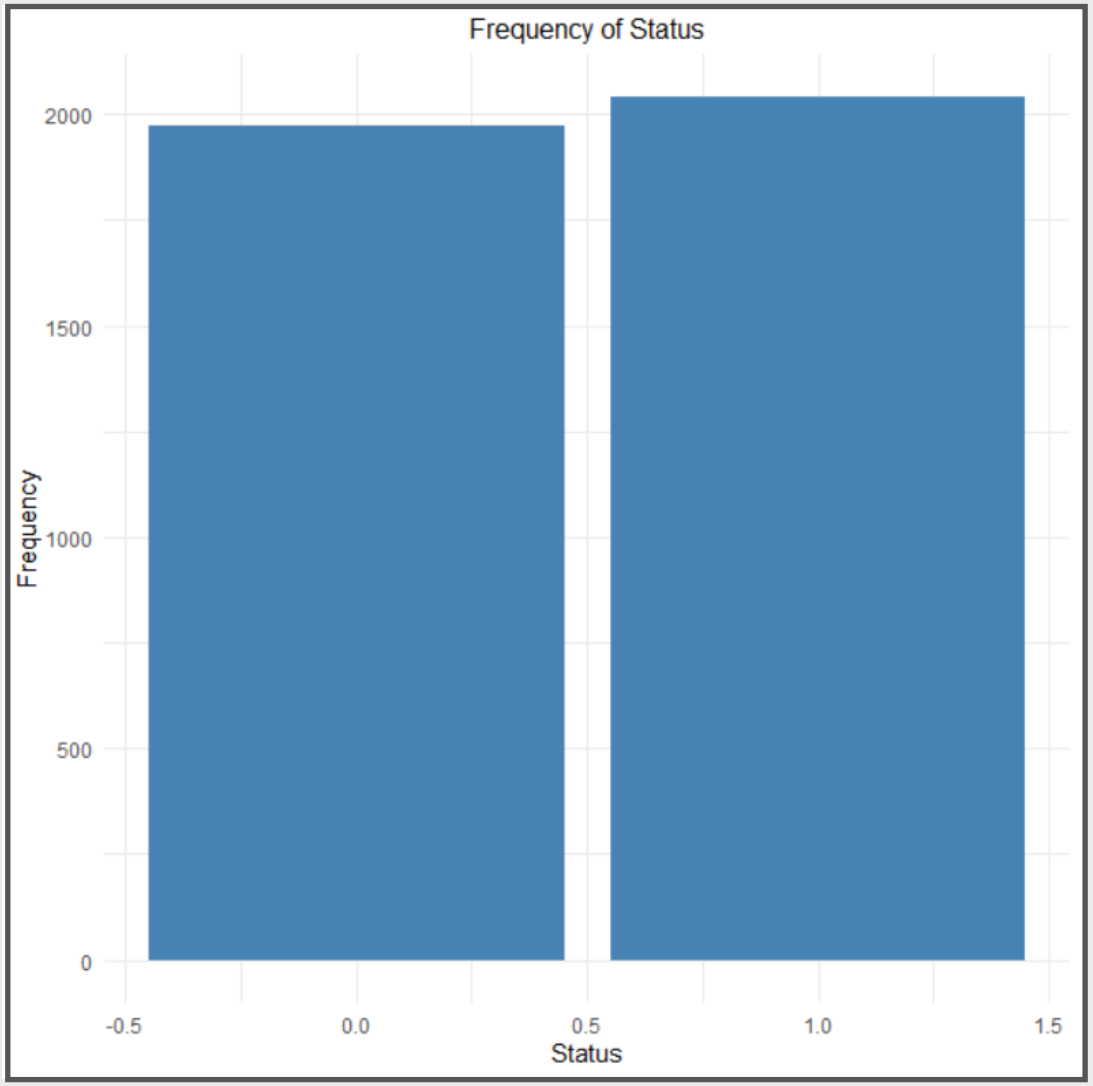
Status가 1인 전복은 성장 가능성이 ↓
➔ 양식을 지속해도 큰 이익을 창출하기 어려움



5. 타깃 변수 범주화 - Status 별 빈도

Frequency of Status	
0	1
1972	2039

범주 별 데이터 불균형이 비교적 해소됨



2 데이터 전처리

6. 변수 선택 – 상관 계수 확인

기존 변수 중 'Sex ','Whole weight', 'Shucked weight', 'Viscera weight', 'Shell weight'

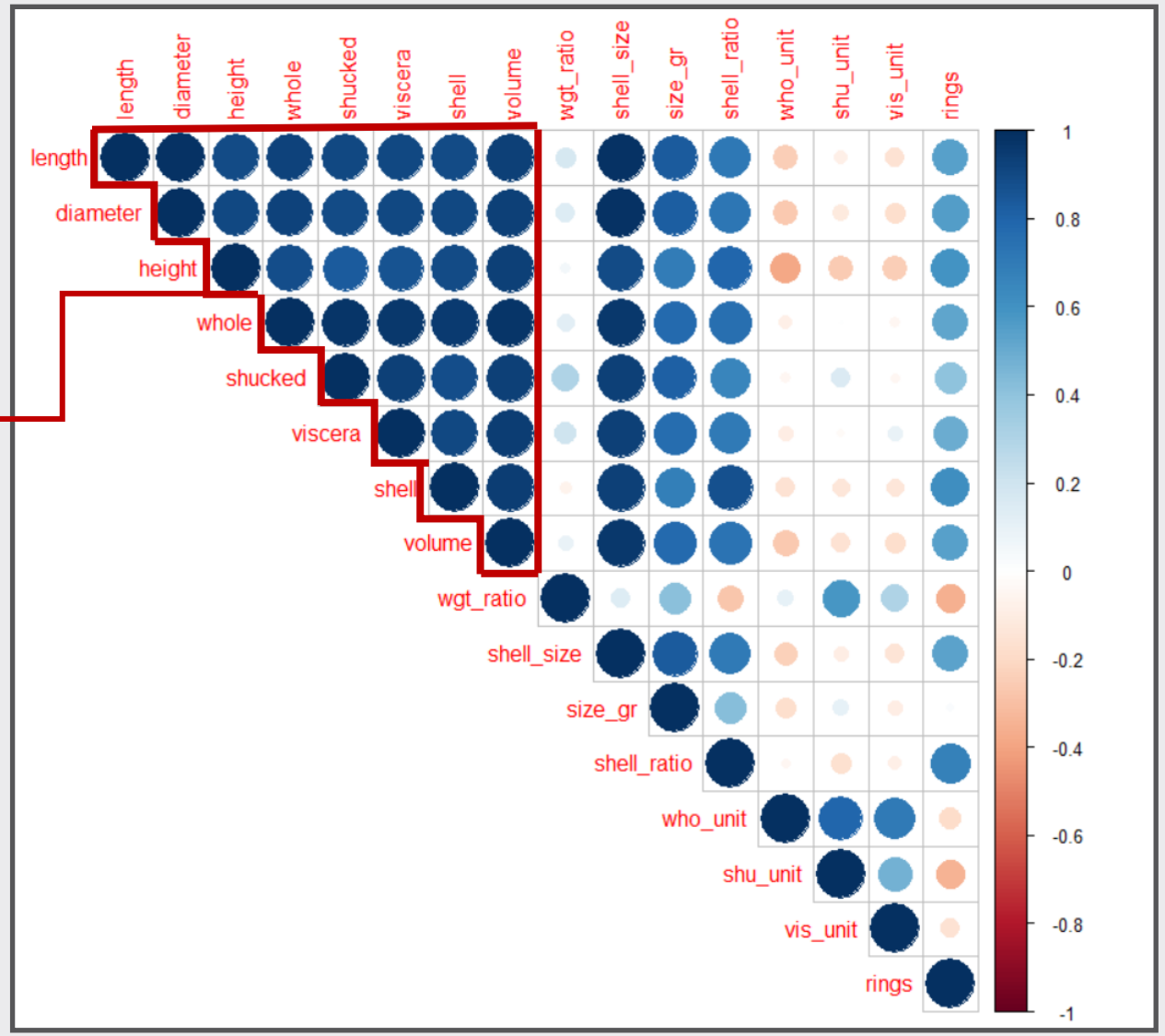
➔ 파생변수들이 그 역할을 대신하고 있음

'Length', 'Diameter'

➔ 상관관계 큼
➔ 'Shell size' 변수 만들었음

'Volume'

➔ 단위 변수들 만드는데 사용하여 제거



2 데이터 전처리

6. 변수 선택 - VIF 확인

	Height	Weight raito	Shell size	Size Growth
VIF	12.141628	44.071348	10.500820	5.049784
	Shell ratio	Whole unit	Shucked unit	Viscera unit
VIF	6.269821	116.675089	101.056129	25.574988

대부분의 변수가 다중 공선성이 매우 큼

전복의 크기에 대한 정보는 Shell size로 알 수 있음
➔ Height 제거

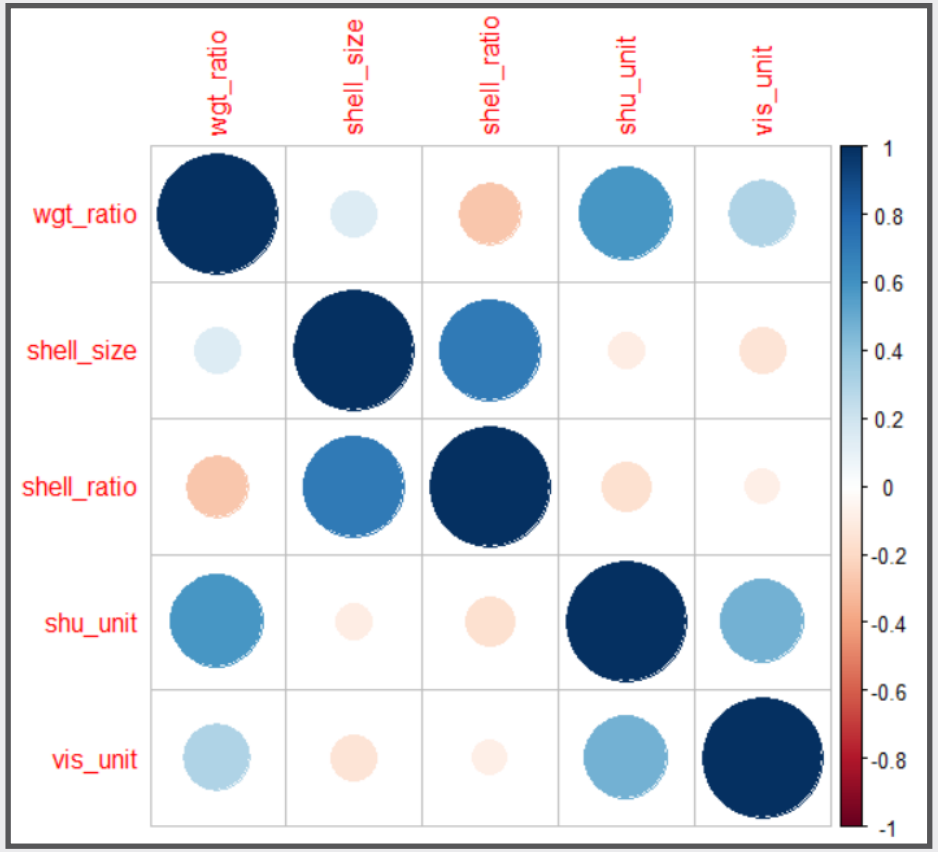
전체 무게에 대한 정보는 살, 내장, 껍질의 무게의 합으로 알 수 있음
➔ Whole unit 제거

Size growth 변수는 rings의 정보를 포함한 변수
➔ Size growth 제거

6. 변수 선택 – 최종 변수

	VIF
(살+내장)무게의 비율	2.4939
껍질의 크기	3.2120
단위 면적당 껍질의 무게	3.1628
단위부피당 살의 무게	1.8951
단위부피당 내장의 무게	1.3620

모든 변수 VIF < 10 임



2 데이터 전처리

7. 데이터 변환 – 정규화

OBS	Adult	Weight Ratio_sc	Shell size_sc	Shell ratio_sc	Shucked unit_sc	Viscera unit _sc	Status
1	1	0.5981346	0.3160797	0.3023258	0.4892487	0.3419855	1
2	1	0.6591272	0.1676786	0.2369367	0.3770095	0.2955172	0
…	…	0.4780570	0.4304797	0.3200239	0.2125549	0.2089937	0
4105	1	…	…	…	…		
4106	1	0.7517565	0.5567699	0.3804880	0.2348570	0.2256609	0

모든 연속형 변수의 값이 0~1 사이


3
모델링

3 모델링

로지스틱 모형 : 오즈 해석 등 모형 해석이 가능함.

랜덤 포레스트 모형 : 이상치에 덜 민감. 변수 중요도 확인

평가 지표

이익 도표는 로지스틱 모형을 평가하는 용도로만 사용하되,

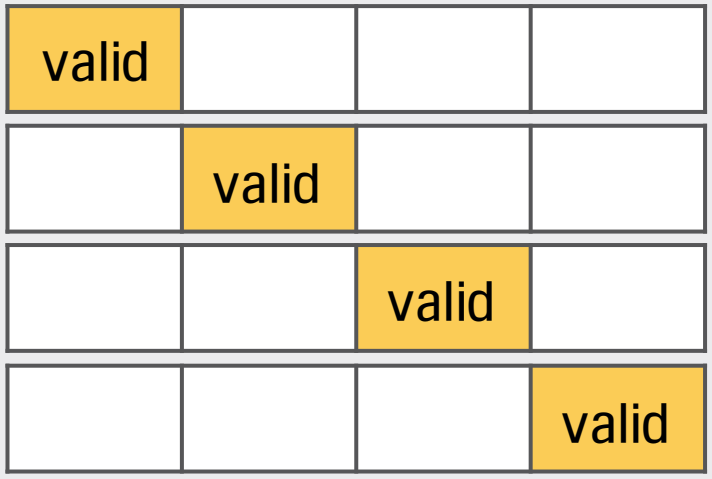
k겹 교차검증으로 정확도를 구하여 두 모형을 비교함.

3 모델링

1. k겹 교차검증을 활용한 로지스틱 회귀 모형 구축

1. train : test = 8 : 2 로 분리

2. 4겹 교차검증



0부터 1까지 0.025 단위로 cutoff 후보 값 만든 후,
평균 정확도가 가장 높았던 최적 cutoff값 도출.

교차 검증 결과	
최적 모형	Full model
최적 Cutoff 값	0.525
평균 정확도	0.7934

3 모델링

1. k겹 교차검증을 활용한 로지스틱 회귀 모형 구축

구축된 모형의 회귀계수		
Intercept	-1.9128	***
Adult	0.6830	***
Weight_ratio_sc	-4.4203	***
Shell_ratio_sc	6.8455	***
Shell_size_sc	5.5460	***
Shucked_unit_sc	-4.3082	***
Viscera_unit_sc	3.3399	***

-Test set 이용-

	실제 범주		
예측 범주	status	0	1
	0	318	70
	1	89	325

정확도	민감도	특이도
80.17%	78.13%	82.28%

3 모델링

2. 구축된 로지스틱 모형 평가

전체 평균에 비해 십분위 0에 1.93배 높은 비율로
status=1 이 포함 되어있음

decile	Predicted Prob	% of Status=1	Cum % of Status=1	# of Status=1	% of Total Status=1	Cum # of Status=1	Cum % of Total Status=1	Lift(%)	Cum Lift(%)
0	97.24%	98.77%	98.77%	80	19.32%	80	19.32%	193.24	193.24
1	89.78%	88.75%	93.79%	71	17.15%	151	36.47%	171.50	182.37
2	81.04%	83.75%	90.46%	67	16.18%	218	52.66%	161.84	175.52
3	71.93%	73.75%	86.29%	59	14.25%	277	66.91%	142.51	167.27
4	58.35%	61.25%	81.30%	49	11.84%	326	78.74%	118.36	157.49
5	45.74%	41.25%	74.64%	33	7.97%	359	86.71%	79.71	144.52
6	30.59%	33.75%	68.81%	27	6.52%	386	93.24%	65.22	133.20
7	18.21%	26.25%	63.49%	21	5.07%	407	98.31%	50.72	122.89
8	8.67%	7.50%	57.28%	6	1.45%	413	99.76%	14.49	110.84
9	2.92%	1.23%	51.62%	1	0.24%	414	100.00%	2.42	-
total	50.45%	51.62%	-	414	100%	-	-	-	-

3 모델링

2. 구축된 로지스틱 모형 평가

십분위 0부터 십분위 4까지 구축된 로지스틱 모형이
평균 모형보다 더 효율적

decile	Predicted Prob	% of Status=1	Cum % of Status=1	# of Status=1	% of Total Status=1	Cum # of Status=1	Cum % of Total Status=1	Lift(%)	Cum Lift(%)
0	97.24%	98.77%	98.77%	80	19.32%	80	19.32%	193.24	193.24
1	89.78%	88.75%	93.79%	71	17.15%	151	36.47%	171.50	182.37
2	81.04%	83.75%	90.46%	67	16.18%	218	52.66%	161.84	175.52
3	71.93%	73.75%	86.29%	59	14.25%	277	66.91%	142.51	167.27
4	58.35%	61.25%	81.30%	49	11.84%	326	78.74%	118.36	157.49
5	45.74%	41.25%	74.64%	33	7.97%	359	86.71%	79.71	144.52
6	30.59%	33.75%	68.81%	27	6.52%	386	93.24%	65.22	133.20
7	18.21%	26.25%	63.49%	21	5.07%	407	98.31%	50.72	122.89
8	8.67%	7.50%	57.28%	6	1.45%	413	99.76%	14.49	110.84
9	2.92%	1.23%	51.62%	1	0.24%	414	100.00%	2.42	-
total	50.45%	51.62%	-	414	100%	-	-	-	-

3 모델링

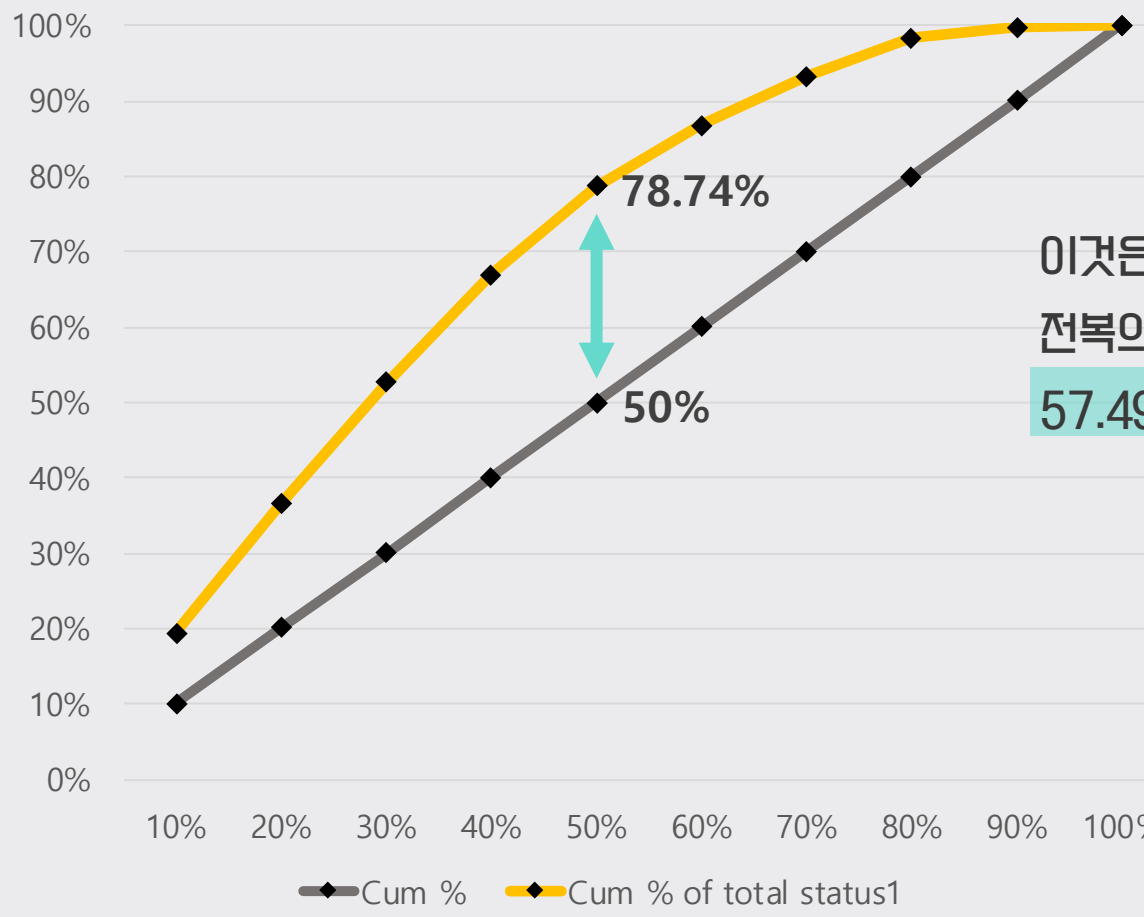
2. 구축된 로지스틱 모형 평가

십분위0 부터 십분위 4까지 전체 데이터의 절반만으로도
Status=1인 전복을 78.74% 찾을 수 있음

decile	Predicted Prob	% of Status=1	Cum % of Status=1	# of Status=1	% of Total Status=1	Cum # of Status=1	Cum % of Total Status=1	Lift(%)	Cum Lift(%)
0	97.24%	98.77%	98.77%	80	19.32%	80	19.32%	193.24	193.24
1	89.78%	88.75%	93.79%	71	17.15%	151	36.47%	171.50	182.37
2	81.04%	83.75%	90.46%	67	16.18%	218	52.66%	161.84	175.52
3	71.93%	73.75%	86.29%	59	14.25%	277	66.91%	142.51	167.27
4	**58.35%**	**61.25%**	**81.30%**	**49**	**11.84%**	**326**	**78.74%**	**118.36**	**157.49**
5	45.74%	41.25%	74.64%	33	7.97%	359	86.71%	79.71	144.52
6	30.59%	33.75%	68.81%	27	6.52%	386	93.24%	65.22	133.20
7	18.21%	26.25%	63.49%	21	5.07%	407	98.31%	50.72	122.89
8	8.67%	7.50%	57.28%	6	1.45%	413	99.76%	14.49	110.84
9	2.92%	1.23%	51.62%	1	0.24%	414	100.00%	2.42	-
total	50.45%	51.62%	-	414	100%	-	-	-	-

3 모델링

2. 구축된 로지스틱 모형 평가

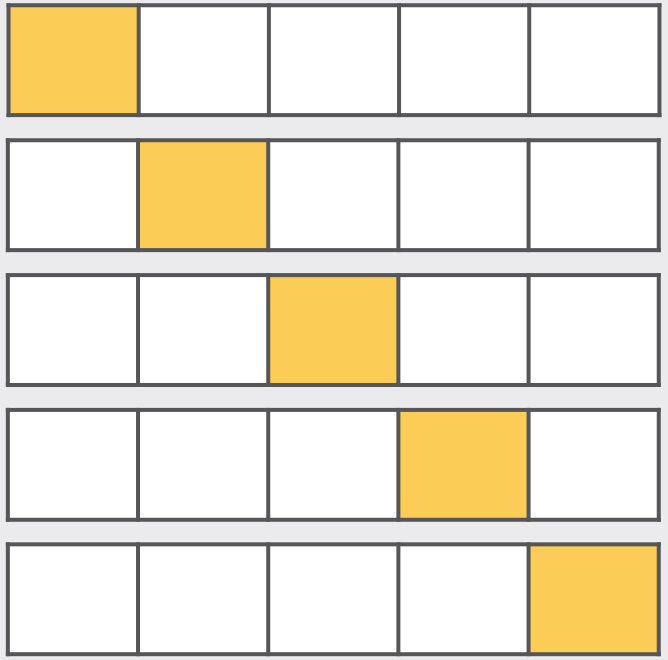


이것은 무작위로 전복을 뽑아
전복의 성장 정도를 예측했을 때보다
57.49% 증가한 수치

3 모델링

3. k겹 교차검증을 활용한 랜덤 포레스트 모형 구축

1. 5겹 교차검증 (train : test = 8 : 2)



교차 검증 결과	
1	0.7783
2	0.8105
3	0.7993
4	0.7930
5	0.7780
평균 정확도	0.7918

평균 정확도와 가장 유사,
이때의 혼돈행렬과
변수 중요도를 도출함

3 모델링

3. 구축된 랜덤 포레스트 모형 평가

-변수 중요도-

-검증 데이터 이용-

	실제 범주		
예측 범주	status	0	1
	0	299	96
	1	71	336

정확도	민감도	특이도
79.30%	81.03%	77.83%

4 결론

4 결론

1. 오즈 해석

성장 정체기 : status=1 , 성장기 : status=0

변수명	성장 정체기 전복의 오즈 변화
성별 정해진 전복	1.98 배

변수명	한 단위 증가 할 때 성장 정체기 전복의 오즈 변화
전체 무게 중 (살+내장) 비율	0.188 배
단위 면적 당 껍질 무게	1.104 배
단위 부피 당 내장의 무게	0.996 배
단위 부피 당 내장의 무게	1.002 배

※ 정규화 전으로 변환하여 오즈 변화 해석함

4 결론

2. 중요 변수 이익 도표 도출

decile	adult	shell_size	weight_ratio	shucked_unit	viscera_unit	shell_ratio
0	98.77%	10008.34	56.09%	0.000407	0.000233	0.008985
1	96.25%	9760.95	62.52%	0.000463	0.000249	0.007640
2	93.75%	9127.77	64.06%	0.000473	0.000246	0.007073
3	91.25%	8627.17	65.25%	0.000493	0.000256	0.006770
4	90.00%	7765.13	65.57%	0.000514	0.000257	0.006427
5	76.25%	7042.99	65.84%	0.000519	0.000262	0.006138
6	66.25%	6204.43	66.10%	0.000517	0.000252	0.005700
7	33.75%	5231.13	65.36%	0.000516	0.000255	0.005218
8	18.75%	4194.96	65.73%	0.000544	0.000252	0.004738
9	6.17%	3014.78	67.39%	0.000600	0.000271	0.003862

-Test set 이용한 중요 변수 이익 도표-

1. 성별이 정해진 전복일수록
2. 껍질 면적이 클수록
3. 전체 무게 중 살과 내장의 비율이 감소할수록
4. 단위 부피당 살의 무게가 작을수록
5. 단위 면적당 껍질 무게가 클수록

성장 정체기 전복일 가능성이 높다.

4 결론

3. 모형의 활용 제안

추가적으로 비용을
투자할 가치가 있음

성장기 전복

VS

성장 정체기 전복

양식을 중단하고
판매하는 것이 더 효율적

전복의 성장 정도 분류 모형 → 양식업자들이 특정 전복을 계속해서 양식할 지 판단할 근거 제공

4 결론

4. 연구의 한계 및 향후 과제

감사합니다!

발표 들어주셔서 감사합니다

Q&A