**데이터마이닝을 통한**

**전복의 성장세 여부 분석**

**통계학과 2018110475 김기호**

**통계학과 2017112282 김현**

**통계학과 2018110476 배정민**

**통계학과 2018110499 양승엽**

**< 목 차 >**

**Ⅰ. 서론**

1. 연구 목적

2. 연구 방법

**Ⅱ. 본론**

1. 데이터 소개

2. 데이터 전처리

1) 변수 단위 변환

2) 비논리적 관측치 제거

3) 파생 변수 생성

4) Rings 변수 범주화

5) 변수선택

6) 데이터 변환

3. 모형 구축 및 평가

1) GirdSearchCV를 활용한 로지스틱 회귀분석

2) GirdSearchCV를 활용한 XGBoost

3) GirdSearchCV를 활용한 LightGBM

4) GirdSearchCV를 활용한 랜덤포레스트

**Ⅲ. 결론**

1. 최종 모형 요약

2. 주요 변수 분석과 인사이트 도출

3. 모형 활용 방안

4. 연구의 한계 및 발전 방안

**Ⅳ. 부록**

1. **서론**

**1. 연구 목적**

복은 전복목, 전복과, 전복 속에 속하는 연체동물이다. 몸길이는 2-30cm이며, 대부분 긴 타원형 모형으로 크기와 모양은 종마다 다양하다. 전복은 넓적한 근육성 발이 있어 바위에 붙어 치설로 식물을 갉아 먹는다. 발은 크고 넓으며, 머리에는 한 쌍의 더듬이와 눈이 있다. 성별은 내장 색깔로 구별가능한데, 보통 황백색이면 수컷이고 녹색이면 암컷으로 구분한다. 전복은 물살의 세기와 먹이에 영향을 받으며, 자연산과 양식에 따른 환경 차이에서도 영향을 받는다.

전복은 초기에는 1년에 2~3cm정도 자라지만 그후 점차 성장속도가 줄어든다. 나이에 따라 성장속도가 변하기 때문에 윤문의 구분이 점차 어려워지고 크기로 나이를 추정한다. 산업적인 면에서 바라볼 때, 전복의 성장세를 파악하기 어렵다는 것은 전복 산업의 세분화된 투자에 대한 리스크를 가져다 준다. 이에 본 조는 ‘해양 자원 사단 타로나 해양 연구소 타스마니아 주 산업 수산부’에서 제공한 데이터를 활용해 전복의 나이에 대한 분석을 시행하고, 구체적으로 전복의 성장세와 하락세를 예측하는 모형을 구축한다. 구축된 모형을 통해 전복 산업의 효율적인 운영과 이윤 창출에 대한 제안을 하고자 한다.

**2. 연구 방법**

전복의 나이와 성장세는 관리되는 환경(먹이 등)과 특수한 지역적 조건에 따라 영향을 많이 받는다. 또한 전복의 나이에 따라 성장의 정도도 다르게 진행된다. 그렇기 때문에 다양한 변수를 함께 처리해야 전복의 나이 혹은 윤문의 수를 예측하기가 수월해질 것이다. 그러나 전복에 크게 영향을 주는 변수가 부족하기에 본 프로젝트에서 주어진 외형만으로 전복의 단순 나이를 예측하는 것은 어렵다고 판단했다. 구체적인 전복의 나이를 예측하기보다는 전복이 성장세를 분류하여 예측을 하는 것이 효율적인 분석이 될 것이라 생각한다. 주어진 변수들의 특징을 고려하여 Rings(운문 수) 변수를 성장세에 따라 범주화한 뒤, 이를 예측하는 로지스틱 회귀모형, 부스팅(xgboost, lgbm,), 배깅(랜덤포레스트) 모형을 구축하고 모형 선택을 진행하고자 한다.

1. **본론**

**1. 데이터 소개**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBS** | **Sex** | **Length** | **Diameter** | **Height** | **Whole weight** | **Shucked weight** | **Viscera weight** | **Shell weight** | **Rings** |
| **1** | M | 0.455 | 0.365 | 0.095 | 0.5140 | 0.2245 | 0.1010 | 0.150 | 15 |
| **2** | M | 0.350 | 0.265 | 0.090 | 0.2255 | 0.0995 | 0.0485 | 0.070 | 7 |
| **3** | F | 0.530 | 0.420 | 0.135 | 0.6770 | 0.2565 | 0.1415 | 0.210 | 9 |
| **4** | M | 0.440 | 0.365 | 0.125 | 0.5160 | 0.2155 | 0.1140 | 0.1550 | 10 |
| **…** | … | … | … | …. | … | … | … | … | … |
| **4174** | M | 0.590 | 0.440 | 0.135 | 0.9660 | 0.4390 | 0.2145 | 0.2605 | 10 |
| **4175** | M | 0.600 | 0.475 | 0.205 | 1.1760 | 0.5255 | 0.2975 | 0.3080 | 9 |
| **4176** | F | 0.625 | 0.485 | 0.150 | 1.0945 | 0.5310 | 0.2610 | 0.2960 | 10 |
| **4177** | M | 0.710 | 0.555 | 0.195 | 1.9485 | 0.9455 | 0.3765 | 0.4950 | 12 |

[ 표 2 - 1 - 1 ] 불러온 데이터

|  |  |  |
| --- | --- | --- |
| **변수명** | **변수 설명** | **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 | grams |
| **Shucked weight** | 내장을 제외한 전복살의 무게 | grams |
| **Viscera weight** | 피를 제거한 내장의 무게 | grams |
| **Shell weight** | 물기를 제거한 껍데기의 무게 | grams |
| **Rings** | 전복의 껍질에 나타난 윤문 수 |  |

[ 표 2 - 1 - 2 ] 데이터 설명

주어진 데이터는 7개의 연속형 변수(Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight)와 1개의 범주형 변수(Sex), 1개의 구간형 변수(Rings)로 총 9개의 변수와 4177개의 행으로 구성되어 있다. 여기서 연속형 변수들의 값들은 모두 200으로 나누어진 값이며, 전복의 길이와 무게에 대한 변수들이다. 길이와 무게에 관련되어 있는 변수들이기 때문에 서로 독립변수들 간의 상관성이 크게 존재함을 확인할 수 있다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Length** | **Diameter** | **Height** | **Whole weight** | **Shucked weight** | **Viscera weight** | **Shell weight** | **Rings** |
| **Length** | 1.0000 | 0.9868 | 0.8276 | 0.9253 | 0.8979 | 0.9030 | 0.8977 | 0.5567 |
| **Diameter** | 0.9868 | 1.0000 | 0.8337 | 0.9255 | 0.8932 | 0.8997 | 0.9053 | 0.5747 |
| **Height** | 0.8276 | 0.8337 | 1.0000 | 0.8192 | 0.7750 | 0.7983 | 0.8173 | 0.5575 |
| **Whole weight** | 0.9253 | 0.9255 | 0.8192 | 1.0000 | 0.9694 | 0.9664 | 0.9554 | 0.5404 |
| **Shucked weight** | 0.8979 | 0.8932 | 0.7750 | 0.9694 | 1.0000 | 0.9320 | 0.8826 | 0.4208 |
| **Viscera weight** | 0.9030 | 0.8997 | 0.7983 | 0.9664 | 0.9320 | 1.0000 | 0.9077 | 0.5038 |
| **Shell weight** | 0.8977 | 0.9053 | 0.8173 | 0.9554 | 0.8826 | 0.9077 | 1.0000 | 0.6276 |
| **Rings** | 0.5567 | 0.5747 | 0.5575 | 0.5404 | 0.4209 | 0.5038 | 0.6276 | 1.0000 |

[ 표 2 - 1 -3 ] 상관계수 행렬

**2. 데이터 전치리**

1) 변수 단위 변환

먼저, 주어진 연속형 변수들은 모두 200으로 나누어진 뒤 제공된 데이터이다. 200으로 나누기 전의 수치들이 직관적인 이해를 돕는다고 판단하여 변수들에 200을 다시 곱하여 변수 단위를 변환시켰다.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBS** | **Sex** | **Length** | **Diameter** | **Height** | **Whole weight** | **Shucked weight** | **Viscera weight** | **Shell weight** | **Rings** |
| **1** | M | 91 | 73 | 19 | 102.8 | 44.9 | 20.2 | 30 | 15 |
| **2** | M | 70 | 53 | 18 | 45.1 | 19.9 | 9.7 | 14 | 7 |
| **3** | F | 106 | 84 | 27 | 135.4 | 51.3 | 28.3 | 42 | 9 |
| **4** | M | 88 | 73 | 25 | 103.2 | 43.1 | 22.8 | 31 | 10 |
| **…** | … | … | … | … | … | … | … | … | … |
| **4174** | M | 118 | 88 | 27 | 193.2 | 87.8 | 42.9 | 52.1 | 10 |
| **4175** | M | 120 | 95 | 41 | 235.2 | 105.1 | 57.5 | 61.6 | 9 |
| **4176** | F | 125 | 97 | 30 | 218.9 | 106.2 | 52.2 | 59.2 | 10 |
| **4177** | M | 142 | 111 | 39 | 389.7 | 189.1 | 75.3 | 99.0 | 12 |

[ 표 2 - 2 - 1 - 1 ] 단위 변환 데이터

2) 이상 관측치 제거

주어진 변수들의 분포와 통계량들을 확인하고 탐색적 자료분석 과정을 거쳤다. 그 과정에서 실제로 관측되기 어렵다고 판단된 이상 관측치들을 찾았고 이를 제거하거나 대체하는 방법으로 처리했다.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBS** | **Sex** | **Length** | **Diameter** | **Height** | **Whole weight** | **Shucked weight** | **Viscera weight** | **Shell weight** | **Rings** |
| **43** | I | 48 | 35 | 9 | 14.0 | 6.3 | 4.7 | 4 | 5 |
| **44** | I | 41 | 30 | 11 | 8.4 | 5.1 | 3 | 2.4 | 5 |
| **45** | I | 42 | 30 | 10 | 8.4 | 3.5 | 2.5 | 3 | 4 |
| **46** | I | 78 | 59 | 19 | 40.6 | 17.5 | 9 | 15 | 7 |
| **…** | … | … | … | … | … | … | … | … | … |
| **3970** | I | 76 | 60 | 18 | 55.4 | 33.1 | 12.5 | 16.4 | 6 |
| **3997** | I | 63 | 46 | 0 | 26.8 | 11.5 | 5.7 | 70.1 | 6 |
| **4047** | M | 120 | 92 | 31 | 133.1 | 57.0 | 29.8 | 53.8 | 11 |
| **4144** | F | 131 | 100 | 44 | 271.8 | 128.4 | 65.1 | 81.0 | 13 |

[ 표 2 - 2 - 2 - 1 ] 무게 이상치

전복에는 단순히 살과 내장, 껍데기뿐만 아니라 수분이나 피 등의 추가적인 물질들로 구성이 되어 있다. 즉, 세가지 변수 살의 무게(Shucked weight)과 내장의 무게(Viscera weight), 껍데기의 무게(Shell weight)를 합한 값이 의미하는 무게는 전체 무게(Whole weight)보다 논리적으로 클 수가 없다. 따라서 이러한 관측치에 대해서는 모두 제거를 하였다(161개).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBS** | **Sex** | **Length** | **Diameter** | **Height** | **Whole weight** | **Shucked weight** | **Viscera weight** | **Shell weight** | **Rings** |
| **1258** | I | 86 | 68 | 0 | 85.3 | 41.3 | 17.2 | 23.0 | 8 |

[ 표 2 - 2 - 2 - 2 ] 두께(Height) 이상치

그 다음으로 전복의 두께(Height)가 0인 하나의 관측치를 확인했다. 전복의 두께 (Height)와 상관계수가 가장 높았던 수직 길이(Diameter)를 활용하여, 같은 수직 길이(Diameter)값을 갖는 모든 관측치들의 긴 부분의 길이(Length) 평균을 사용하여 대체했다(1개).

|  |  |
| --- | --- |
|  |  |

[ 그림 2 - 2 - 2 - 1 ] 두께 박스플랏, 히스토그램 [ 그림 2 - 2 - 2 - 2] 두께와 전체 무게 그래프

위와 같이 모든 변수들의 분포를 확인하는 과정에서 [ 그림 2 - 2 - 2 - 2] 상의 3개의 이상 의심 관측치가 관측됐다. 이 관측치들을 구체적으로 탐색해보았다.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBS** | **Sex** | **Length** | **Diameter** | **Height** | **Whole weight** | **Shucked weight** | **Viscera weight** | **Shell weight** | **Rings** |
| **1418** | M | 141 | 113 | 103 | 442 | 221.5 | 97.3 | 102.4 | 10 |
| **2052** | F | 91 | 71 | 226 | 118.8 | 66.4 | 23.2 | 26.7 | 8 |

[ 표 2 - 2 - 2 - 3 ] 극단적으로 큰 값을 갖는 두께(Height)

그 중 두개의 관측치는 전체 무게(Whole weight)에 비해 두께(Height)가 극단적으로 크기 때문에 논리적이지 않다고 판단하여 제거했다(2개).

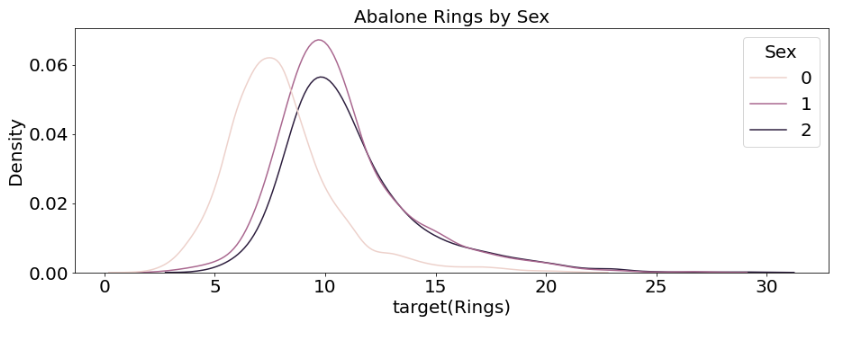
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBS** | **Sex** | **Length** | **Diameter** | **Height** | **Whole weight** | **Shucked weight** | **Viscera weight** | **Shell weight** | **Rings** |
| **1175** | F | 127 | 99 | 3 | 231.3 | 102.3 | 61.6 | 57.7 | 9 |
| **2170** | I | 33 | 23 | 3 | 2.9 | 1.1 | 0.6 | 1 | 2.7 |

[ 표 2 - 2 - 2 - 4 ] 두께(Height) 3이하인 관측치

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **25** | F | 123 | 96 | 33 | 232.3 | 102.6 | 60.27 | 61 | 10 |
| **105** | M | 121 | 94 | 32 | 234.7 | 99.5 | 48.1 | 69 | 12 |
| **…** | … | … | … | … | … | … | … | … | … |
| **1175** | F | 127 | 99 | 3 | 231.3 | 102.3 | 61.6 | 57.7 | 9 |
| **…** | … | … | … | … | … | … | … | … | … |
| **4056** | F | 129 | 100 | 30 | 231.8 | 93.5 | 67.1 | 62 | 9 |

[ 표 2 - 2 - 2 - 5 ] 비슷한 전체 무게(Whole weight)를 기준으로 살핀 두께(Height) 이상치

[ 표 2 - 2 - 2 - 4 ]를 보고 두께(Height)가 매우 작으면서 무게 관련 변수들은 높은 관측치가 이상치로 의심되었다. 이에 비슷한 전체 무게(Whole weight) 그룹을 갖는 관측치들을 모두 비교하면서 1175행의 두께(Height)가 무게를 고려할 때 논리적으로 나올 수 없다고 판단했고 이를 제거했다(1개).



[ 그림 2 - 2 - 2 - 3] 두께와 전체 무게 그래프

추가적으로 성별(Sex)의 범주 별 특징을 분포상에서 파악하여 재범주화하였다. 치패(I)는 암(F),수(M)와 다른 양상을 띄고 있음을 확인했고 치패는 0으로 암과 수는 성패(1)로 재범주화하여 Adult 변수를 만들었다.

3) 파생 변수 생성

앞서 데이터 소개 과정에서 보았듯이 [ 표 2 - 1 -3 ] 상관계수 행렬을 보면 독립변수들 간의 상관계수가 매우 높음을 확인할 수 있다. 이에 다중공선성이 의심되어 VIF값들을 확인해보았다.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Length** | **Diameter** | **Height** | **Whole weight** | **Shucked weight** | **Viscera weight** | **Shell weight** |
| **Vif** | 41.690171 | 42.754963 | 6.774577 | 143.285733 | 35.667654 | 19.420346 | 26.192182 |

[ 표 2 - 2 - 3 - 1 ] 연속형 변수들의 Vif 값

[ 표 2 - 2 - 2 - 5 ]를 보면 Height을 제외한 모든 변수의 Vif값이 10보다 큰 값임을 확인할 수 있다. 이는 본 데이터가 전복의 무게 관련 변수와 길이 및 크기 관련 변수로 구성되어 있기 때문이며, 다중공선성 문제가 발생할 수 밖에 없는 변수 구성이다. 다중공선성이 존재하면 모형을 추정하기 어렵기 때문에 다중공선성 제거를 위해 기존 변수들을 활용한 파생변수 생성을 고민하였다. 크기와 관련이 있는 부피, 수평면적 그리고 무게와 관련이 있는 수분 포함량, 밀도를 생성했다.

|  |  |  |
| --- | --- | --- |
| **변수명** | **변수 설명** | **계산식** |
| **Volume** | 전복의 부피(반구 부피 기준) | π |
| **Moisture** | 전복의 수분 포함량 |  |
| **Area** | 전복의 수평 단면적 |  |
| **Density** | 전복의 단위 부피당 무게(밀도) |  |

[ 표 2 - 2 - 3 - 2 ] 파생변수 계산식

생성한 파생 변수들의 분포는 다음과 같다.

|  |  |
| --- | --- |
|  |  |

[ 그림 2 - 2 - 3 - 1] 부피 박스플랏, 히스토그램 [ 그림 2 - 2 - 3 - 2] 수분 박스플랏, 히스토그램

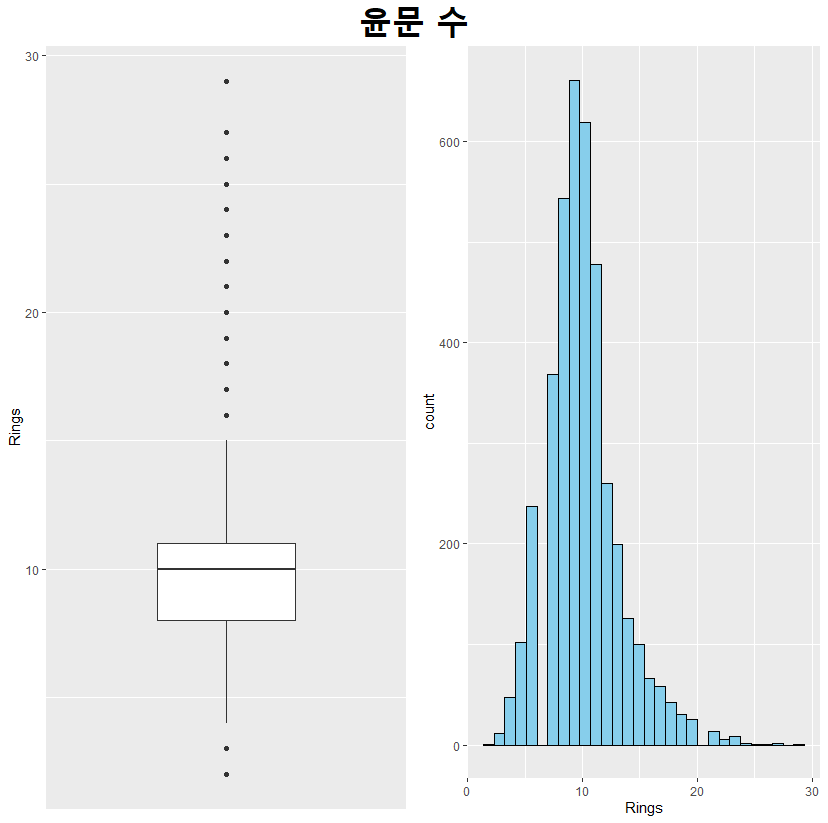
부피 박스플랏에서 이상치는 큰 값 쪽에서 형성되었으나, 크기가 크거나 무게가 많이 나가는 관측치들이 큰 부피 값을 보이는 것으로 확인되어 제거하지 않았다. 수분 역시 이상치가 많이 존재하는 것처럼 보이나 자료의 불균형이 심하고 윤문(Rings) 등 다른 변수 값이 클 때 수분량도 많은 경우가 있는 것으로 확인되어 제거하지 않았다.

|  |  |
| --- | --- |
|  |  |

[ 그림 2 - 2 - 3 - 3] 면적 박스플랏, 히스토그램 [ 그림 2 - 2 - 3 - 4] 밀도 박스플랏, 히스토그램

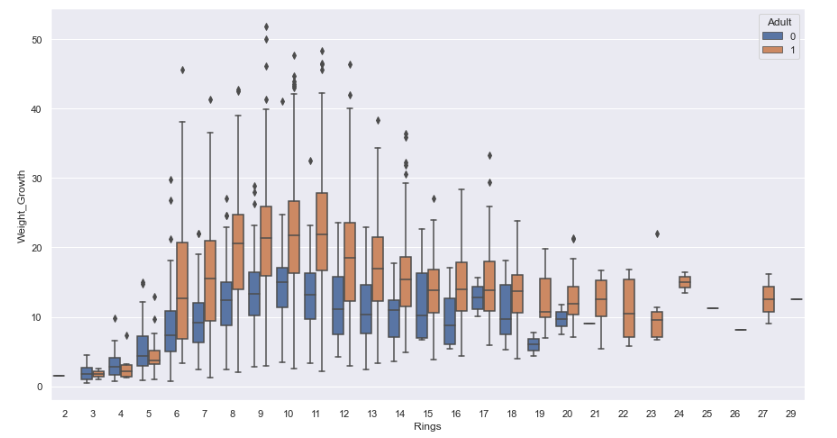
이 두 변수에서 이상치로 눈에 띄는 상위 두 개의 관측치는 같은 관측치로 다른 전복들에 비해 극단적으로 큰 전복으로 확인되었다. 따라서 일반적인 전복이 아니라고 판단해 제거하였다(2개).

4) Rings 변수 범주화



[ 그림 2 - 2 - 3 - 5] 윤문 수(Rings)의 박스플랏, 히스토그램

먼저, Rings는 1~29까지 지나치게 세분화되어 있으며 특정 구간에 대부분의 빈도가 밀집되어 있어 양 끝 값의 Rings의 특징을 파악하고 분석하기가 어렵다. 또한 초반에 언급했듯이 주어진 전복의 크기와 무게 및 물리적인 특징들은 먹이와 환경 요소 등에 큰 영향을 받기 때문에 주어진 물리적 특성의 데이터만으로는 세분화된 윤문 수(Rings)를 예측하기가 어렵다. 따라서 전복의 전체 무게 평균 성장률 ()을 만들어 전복의 성장세를 파악하고 이를 통해 Rings를 재범주화하고자 했다.



[ 그림 2 - 2 - 3 - 6] Rings에 대한 Weight growth

[ 그림 2 - 2 - 3 - 6]를 보면 Rings의 값 11까지는 전복의 전체 무게 평균 성장률이 상승세이지만, 그 이후부터는 성장률이 하락세임을 확인할 수 있었다. 따라서 이를 통해 전복의 성장률이 상승세(0)인지 하락세(1)인지를 구분하는 범주형 변수 Status로 만들었다.

5) 변수 선택

기존 변수들은 [ 표 2 - 2 - 3 - 1 ]에서 살핀 바와 같이 다중공선성의 위험이 컸다. 새로 생성한 변수들 역시 기존 변수로부터 파생되었기 때문에 다중공선성 문제를 고려해야 한다. 따라서 앞선 방식과 마찬가지로 다중공선성을 확인한 뒤, 의미가 겹치는 변수들을 제거함으로써 범주형 변수인 성장세(Status)와 연속형 변수 부피(Volume), 수분포함량(Moisture), 밀도(Density)를 남겼다. 이 과정은 연속형 변수들을 통해 범주형 변수 성장세(Status)를 예측하는 모형을 구축하기 위한 사전 작업이다.

6) 데이터 변환

변수마다 단위가 다르기 때문에 변수 선택에 이어서 데이터 변환(scaling)을 진행했다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **OBS** | **Status** | **Volume** | **Moisture** | **Density** |
| **1** | 1 | -0.934644 | -0.283125 | 2.088153 |
| **2** | 0 | -1.293239 | -0.861550 | 0.639543 |
| **3** | 0 | -0.245708 | 0.285971 | -0.528174 |
| **4** | 0 | -0.727205 | -0.413737 | 0.298404 |
| **…** | … | … | … | … |
| **4007** | 0 | -0.004622 | -0.031230 | 0.782640 |
| **4008** | 0 | 1.123778 | 0.024747 | -1.152941 |
| **4009** | 0 | 0.498438 | -0.880209 | -0.126473 |
| **4010** | 1 | 2.012576 | 1.452151 | 0.208426 |

[ 표 2 - 2 - 6- 1] 파생변수 계산식

**3. 모형 구축 및 평가**

본 조는 전복의 판매 가능성 판단을 위해 로지스틱 회귀, XGBoost, LightGBM, 랜덤포레스트 총 4개의 모형을 구축하고자 한다. 구축한 네 가지의 모형을 평가하기 위해 정확도(Accuracy), 민감도(Sensitivity), 특이도(Specificity), 양성예측도(Positive Predictive Value), 음성예측도(Negative Predictive Value), f1-Score 6가지 지표를 사용할 것이다.

모델 평가를 위해 사용하는 지표들은 단 한번의 데이터셋 분리를 통해 구해지는 값이다. 모델 구축에 필요한 데이터셋 분리 과정을 2번 이상 거쳐 더 정밀한 정확도(Accuracy)를 구하기 위해 그리드서치 교차 검증(Grid Search Cross Validation)기법을 활용해 모델을 평가할 것이다. 각 모델 별 구축 및 평가는 이어서 자세히 설명하도록 한다.

**1) 그리드서치 교차 검증을 활용한 로지스틱 회귀모형 구축**

본 조는 그리드서치 교차 검증을 통해 로지스틱 회귀모형을 구축하기위해 모형 구축에 쓰일 데이터를 Train / Test set의 두가지로 분리하였다. 이때, 각각의 비율을 80%, 20%로 맞추기 위해 5겹의 교차 검증을 시행하도록 한다.

Train set은 로지스틱 모형구축에 사용되고 test set을 통해 로지스틱 모형의 정확도를 가장 높여줄 수 있는 최적의 cut off 값과 최적의 모형을 찾는다. 그 후 최적의 모형과 cut off 값을 test set에 적용해 모델을 평가하도록 한다.

**① 모형 구축 결과**

본 조가 구축한 모형의 독립변수의 회귀계수는 다음과 같다.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Full model** | | | | | |
|  | **Intercept** | **Volume** | **Adult** | **Moisture** | **Density** |
|  | -1.3870 | 0.3967 | 0.8245 | 0.9033 | -0.3772 |
|  | \*\*\* | \*\*\* | \*\*\* | \*\*\* | \*\*\* |

[ 표 2 – 3 - 1 – 1 ] 로지스틱 회귀모형 독립변수의 회귀계수

본 조가 구축한 Full model의 모든 변수가 유의함을 확인할 수 있다. 이때 각 독립변수의 회귀 계수에 대한 해석은 다음과 같다. Volume이 한 단위 증가할 때, 범주1일 오즈는배 증가한다. 범주1일 오즈는 adult인 전복이 infant인 전복에 비해 128.1% 정도 증가한다. Moisture가 한 단위 증가할 때, 범주1일 오즈는 배 증가한다. Density가 한 단위 증가할 때, 범주1일 오즈는 배 증가한다.

**② 그리드서치 교차 검증 결과**

그리드서치 교차 검증을 통해 찾은 최적의 모델과 최적의 cut off값에 따른 test set의 정확도는 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Hyperparameter** | **Cut off Value** | **Accuracy** |
| Full model | C=1, max\_iter=50 | 0.246 | 0.6999 |

[ 표 2 - 3 – 1 -2 ] 로지스틱 회귀모형 그리드서치 교차 검증 결과

그리드서치 교차 검증 결과, 본 조의 로지스틱 회귀모형에 적합한 최적의 cut off 값은 0.246임을 확인할 수 있다.

**③ Test set을 통한 모델평가**

3-1)-②의 결과를 바탕으로 test set에 모형을 테스트한 결과는 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| **Full model** | | | |
|  | **실제 범주** | | |
| **예측**  **범주** | **범주** | **0** | **1** |
| **0** | 296 | 190 |
| **1** | 51 | 266 |
| **평가지표** | | | |
| **민감도** | | 0.8391 | |
| **특이도** | | 0.6091 | |
| **양성 예측도** | | 0.5833 | |
| **음성 예측도** | | 0.8530 | |
| **정확도** | | 0.6999 | |
| **f1-Score** | | 0.6882 | |

[ 표 2 - 3 – 1 - 3 ] 로지스틱 회귀모형 모델 평가 결과

모형의 ROC 곡선은 부록 <C> [ 그림 2 - 3 – 1 - 1 ]를 참고하도록 한다.

Full model의 경우 정확도가 69.99% 이고 f1-Score가 68.82%의 값을 가지므로 구축된 모형이 전복의 성장세 여부를 준수하게 예측한다고 판단할 수 있다.

**2) 그리드서치 교차 검증을 활용한 XGBoost 모형 구축**

본 조는 그리드서치 교차 검증을 통해 XGBoost 모형을 구축하기위해 모형 구축에 쓰일 데이터를 Train / Test set의 두가지로 분리하였다. 이때, 각각의 비율을 80%, 20%로 맞추기 위해 5겹의 교차 검증을 시행하도록 한다.

Train set은 XGBoost 모형 구축에 사용되고, Test set은 XGBoost 모형을 평가하기 위해 사용된다. 이때, 5번의 서로 다른 Train / Test set분리를 통해 계산된 5개의 정확도 중 가장 높은 정확도를 본 조가 구축한 XGBoost 모형의 정확도로 삼는다.

**① 그리드서치 교차 검증 결과**

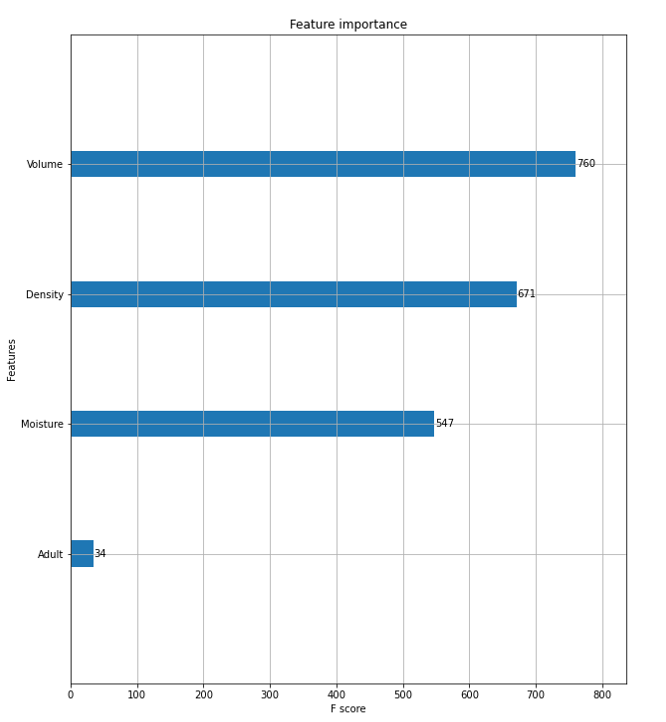
본 조가 구축한 XGBoost 모형에 대한 그리드서치 교차 검증의 결과는 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Hyperparameter** | **Cut off Value** | **Accuracy** |
| Full model | Max\_depth=10, n\_estimator=100, min\_child\_weight=10 | 0.307 | 0.7491 |

[ 표 2 – 3 – 2 – 1 ] XGBoost 모형 그리드서치 교차 검증 결과

그리드서치 교차 검증 결과, 최고 정확도는 0.7491로 가장 높은 정확도를 도출할 수 있으므로 해당 모델의 변수중요도와 혼돈행렬을 참고하도록 한다.

XGBoost 모형의 변수 중요도는 다음과 같다.



[ 그림 2 – 3 – 2 – 1 ] XGBoost 모형 변수 중요도

위 그래프를 통해 XGBoost를 통한 전복의 성장세 여부 예측에 가장 큰 영향을 미치는 상위 3가지 변수는 Volume, Density, Moisture임을 확인할 수 있다.

**② Test set을 통한 모델평가**

3-2)-①의 결과를 바탕으로 test set에 각각의 모형을 테스트한 결과는 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| **XGBoost** | | | |
|  | **실제 범주** | | |
| **예측**  **범주** | **범주** | **0** | **1** |
| **0** | 340 | 146 |
| **1** | 81 | 236 |
| **평가지표** | | | |
| **민감도** | | 0.7445 | |
| **특이도** | | 0.6996 | |
| **양성 예측도** | | 0.6178 | |
| **음성 예측도** | | 0.8076 | |
| **정확도** | | 0.7173 | |
| **f1-Score** | | 0.6853 | |

[ 표 2 – 3 – 2 – 2 ] XGBoost 모델 평가

XGBoost 모형의 경우 정확도가 71.73% 이고 f1-Score가 68.53%의 값을 가지므로 구축된 모형이 전복의 성장세 여부를 준수하게 예측한다고 판단할 수 있다.

**3) 그리드서치 교차 검증을 활용한 LightGBM 모형 구축**

본 조는 그리드서치 교차 검증을 통해 LightGBM 모형을 구축하기위해 모형 구축에 쓰일 데이터를 Train / Test set의 두가지로 분리하였다. 이때, 각각의 비율을 80%, 20%로 맞추기 위해 5겹의 교차 검증을 시행하도록 한다.

Train set은 LightGBM 모형 구축에 사용되고, Test set은 LightGBM 모형을 평가하기 위해 사용된다. 이때, 5번의 서로 다른 Train / Test set분리를 통해 계산된 5개의 정확도 중 가장 높은 정확도를 본 조가 구축한 LightGBM 모형의 정확도로 삼는다.

**① 그리드서치 교차 검증 결과**

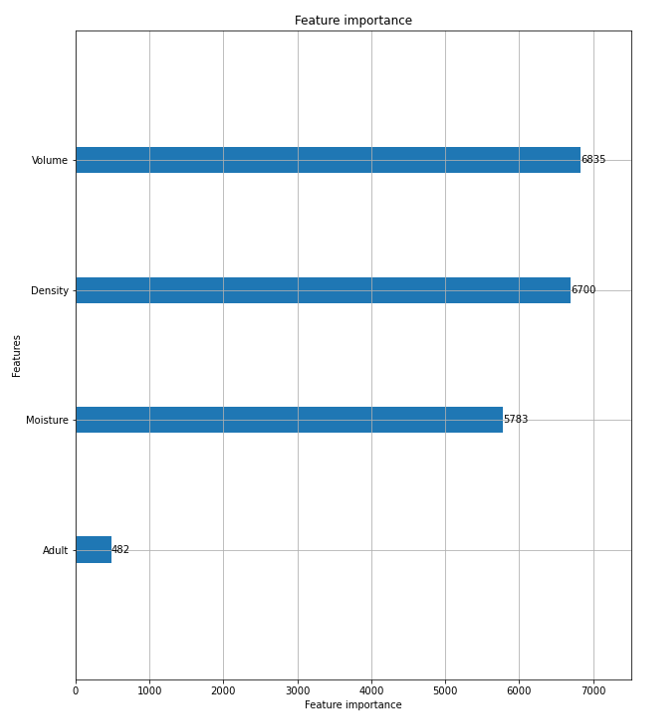
본 조가 구축한 LightGBM 모형에 대한 그리드서치 교차 검증의 결과는 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Hyperparameter** | **Cut off Value** | **Accuracy** |
| Full model | Max\_depth=25, n\_estimator=200, num\_leaves=10,  Learning\_rate=0.01 | 0.343 | 0.7504 |

[ 표 2 – 3 – 3 – 1 ] LightGBM 모형 그리드서치 교차 검증 결과

그리드서치 교차 검증 결과, 최고 정확도는 0.7504로 가장 높은 정확도를 도출할 수 있으므로 해당 모델의 변수중요도와 혼돈행렬을 참고하도록 한다.

LightGBM 모형의 변수 중요도는 다음과 같다.



[ 그림 2 – 3 – 3 – 1 ] LightGBM 모형 변수 중요도

위 그래프를 통해 LightGBM을 통한 전복의 성장세 여부 예측에 가장 큰 영향을 미치는 상위 3가지 변수는 Volume, Density, Moisture임을 확인할 수 있다.

**② Test set을 통한 모델평가**

3-3)-①의 결과를 바탕으로 test set에 각각의 모형을 테스트한 결과는 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| **LightGBM** | | | |
|  | **실제 범주** | | |
| **예측**  **범주** | **범주** | **0** | **1** |
| **0** | 352 | 134 |
| **1** | 97 | 220 |
| **평가지표** | | | |
| **민감도** | | 0.6940 | |
| **특이도** | | 0.7243 | |
| **양성 예측도** | | 0.6215 | |
| **음성 예측도** | | 0.7840 | |
| **정확도** | | 0.7123 | |
| **f1-Score** | | 0.6557 | |

[ 표 2 – 3 – 3 – 2 ] LightGBM 모델 평가

LightGBM 모형의 경우 정확도가 71.23% 이고 f1-Score가 65.57%의 값을 가지므로 구축된 모형이 전복의 성장세 여부를 준수하게 예측한다고 판단할 수 있다.

**4) 그리드서치 교차 검증을 활용한 랜덤포레스트 모형 구축**

본 조는 그리드서치 교차 검증을 통해 랜덤포레스트 모형을 구축하기위해 모형 구축에 쓰일 데이터를 Train / Test set의 두가지로 분리하였다. 이때, 각각의 비율을 80%, 20%로 맞추기 위해 5겹의 교차 검증을 시행하도록 한다.

Train set은 랜덤포레스트 모형 구축에 사용되고, Test set은 랜덤포레스트 모형을 평가하기 위해 사용된다. 이때, 5번의 서로 다른 Train / Test set분리를 통해 계산된 5개의 정확도 중 가장 높은 정확도를 본 조가 구축한 랜덤포레스트 모형의 정확도로 삼는다.

**① 그리드서치 교차 검증 결과**

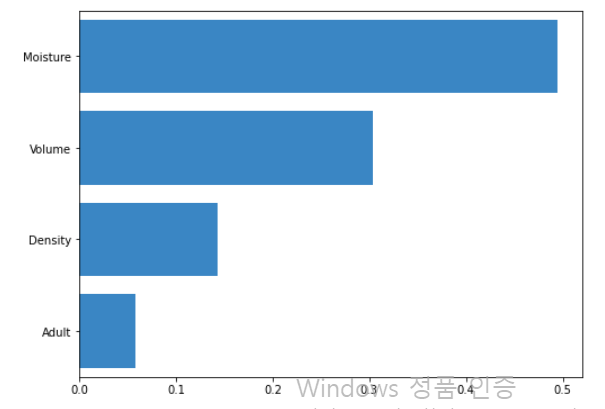
본 조가 구축한 랜덤포레스트 모형에 대한 그리드서치 교차 검증의 결과는 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Hyperparameter** | **Cut off Value** | **Accuracy** |
| Full model | Max\_depth=10, n\_estimator=100, min\_samples\_leaf=10 | 0.344 | 0.7610 |

[ 표 2 – 3 –4 – 1 ] 랜덤포레스트 모형 그리드서치 교차 검증 결과

그리드서치 교차 검증 결과, 최고 정확도는 0.7610로 가장 높은 정확도를 도출할 수 있으므로 해당 모델의 변수중요도와 혼돈행렬을 참고하도록 한다.

랜덤포레스트 모형의 변수 중요도는 다음과 같다.



[ 그림 2 – 3 – 4 – 1 ] 랜덤포레스트 모형 변수 중요도

위 그래프를 통해 랜덤포레스트를 통한 전복의 성장세 여부 예측에 가장 큰 영향을 미치는 상위 3가지 변수는 Moisture, Volume, Density임을 확인할 수 있다.

**② Test set을 통한 모델평가**

3-4)-①의 결과를 바탕으로 test set에 각각의 모형을 테스트한 결과는 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| **랜덤포레스트** | | | |
|  | **실제 범주** | | |
| **예측**  **범주** | **범주** | **0** | **1** |
| **0** | 336 | 150 |
| **1** | 73 | 244 |
| **평가지표** | | | |
| **민감도** | | 0.7697 | |
| **특이도** | | 0.6914 | |
| **양성 예측도** | | 0.6193 | |
| **음성 예측도** | | 0.8215 | |
| **정확도** | | 0.7223 | |
| **f1-Score** | | 0.6864 | |

[ 표 2 – 3 – 4 – 2 ] 랜덤포레스트 모델 평가

랜덤포레스트 모형의 경우 정확도가 72.23% 이고 f1-Score가 68.64%의 값을 가지므로 구축된 모형이 전복의 성장세 여부를 준수하게 예측한다고 판단할 수 있다.

**Ⅲ. 결론**

**1. 최종 모형 요약**

전복 성장세 예측 모형은 전복의 나이 지표인 윤문 수를 예측하기 어렵다는 상황에서 시작하였다. 본 조는 탐색적 자료분석을 진행하며 유사한 윤문 수의 전복 집단에서 무게와 길이 등의 물리적 특성이 다양하게 분포하고 있음을 확인하였고, 자료 조사를 통해서도 윤문 수가 전복의 무게와 길이를 결정하는 절대적인 요소가 아니었음을 확인했다. 따라서 전복의 윤문 수 변수를 범주화 하여 성장세가 상승세인 그룹과 성장세가 하락세인 그룹으로 분류하는 모형을 구축하였다. 로지스틱 모형, XGBoost 모형, LightGBM 모형, 랜덤 포레스트 모형의 정확도가 각각 69.99%, 71.73%, 71.23%, 72.23%로 네 모형 모두 정확도 면에서 지속적으로 성장중인 전복과 성장 속도가 느려진 전복을 준수하게 구분했다.

**2. 주요 변수 분석과 인사이트 도출**

XGBoost와 LightGBM 모형의 변수 중요도 평가에 의하면 부피와 밀도 그리고 수분의 무게 순서로 변수가 중요했으며 랜덤포레스트 모형의 변수 중요도 평가에 의하면 수분의 무게와 부피 그리고 밀도 순서로 중요했다.

앞서 진행한 탐색적 자료분석에 따르면, 성장세가 하락세인 전복(status=1)은 성장세가 상승세인 전복(status=0)에 비해 부피가 큰 경향이 있었고, 반대로 밀도는 작은 경향이 있었다.

앞서 구축한 로지스틱 회귀모형으로 중요 변수에 대한 회귀 계수를 확인한 결과 오즈에 대한 해석을 통해 다음과 같은 인사이트를 도출 할 수 있었다.

1) 부피가 클수록 성장세가 하락세인 전복일 가능성이 높다.

2) 성별이 결정된 전복일수록 성장세가 하락세인 전복일 가능성이 높다.

3) 수분 함유량이 클수록 성장세가 하락세인 전복일 가능성이 높다.

4) 단위 면적 당 전복의 무게가 적을수록 성장세가 하락세인 전복일 가능성이 높다.

**3. 모형 활용 방안**

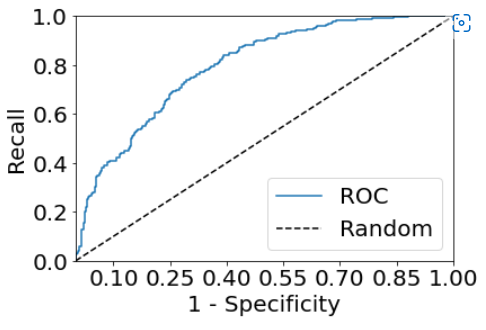
전복 양식 사업장의 목적은 가성비 좋은 전복을 생산해 내는 것이다. 따라서 양식 업자는 각각의 전복들을 계속해서 양식할 가치가 있는지 주기적으로 판단해야 한다. 성장세가 상승세로 판단되는 전복은 추가적으로 시간을 투자할 가치가 있지만, 성장세가 하락세로 판단되는 전복은 양식을 중단하고 판매를 하는 것이 더 효율적인 선택일 수 있다. 본 조가 구축한 모형은 성장세가 상승세인 전복과 성장세가 하락세인 전복의 특성의 차이를 설명하고 있으며 이는 양식업자들이 특정 전복을 계속해서 양식할 지 판단할 근거를 제공한다.

**4. 연구의 한계 및 발전 방안**

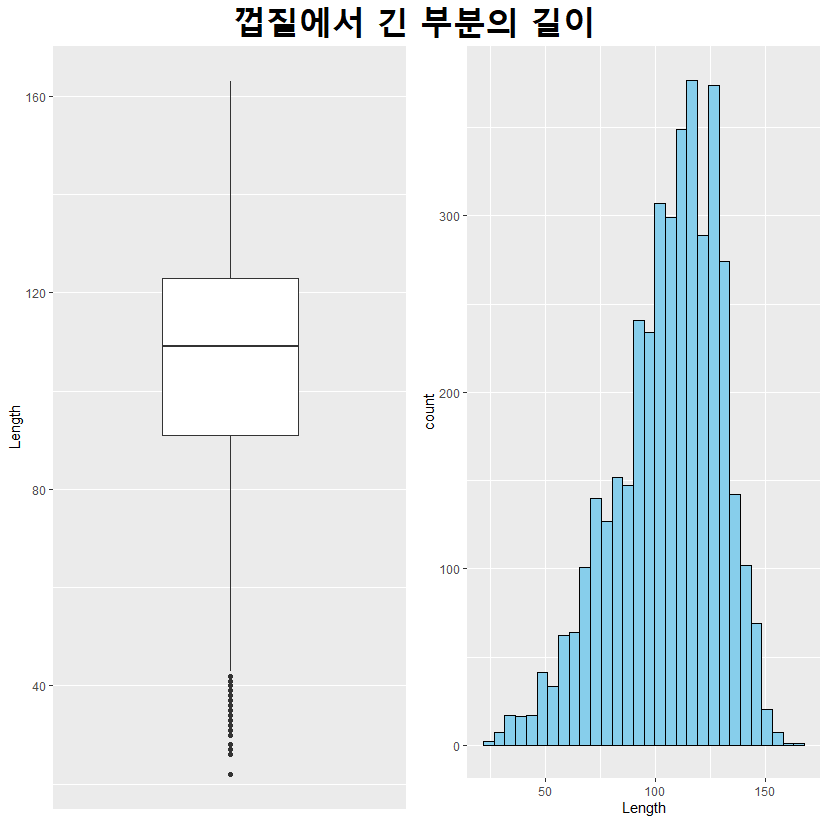
주어진 데이터에는 사실상 유사한 지표를 따라가는 변수들로 구성되어 있었기에 한정된 정보들로 분석을 하여야 했다. 하지만 한층 더 발전된 분석을 위해 탐색적 자료 분석을 진행하여 새로운 변수들을 생성하여 진행하였다. 예시로, 전복 내의 수분 함유량이 윤문 수에 성별이 영향을 끼칠 것이라 생각하여 전체 무게에서 건조된 전복의 무게를 빼주어 Moisture 변수를 만들어주었다. 이처럼 파생변수들을 생성해 윤문 수를 예측하였으나 설명력이 매우 낮게 나와 유의미한 인사이트를 도출해 내는데 한계가 있었고, 이에 따라 연도별 성장세를 이용하여 양식의 효율성에 도움을 주는 모형으로 분석을 진행하였다.

주어진 변수들이 서로 상관성이 매우 높아 사실상 하나의 지표를 따라가는 경향이 있었다는 것이 아쉽다. 또한 전복의 물리적인 특성만 알 수 있다는 점이 아쉬운 점이었다. 전복의 먹이와 서식 환경은 전복의 크기를 결정하는 중요한 요인이다. 따라서 서식 환경에 대해 더 많은 정보가 있다면, 보다 설명력 높은 모형을 구축할 수 있을 것으로 기대된다.

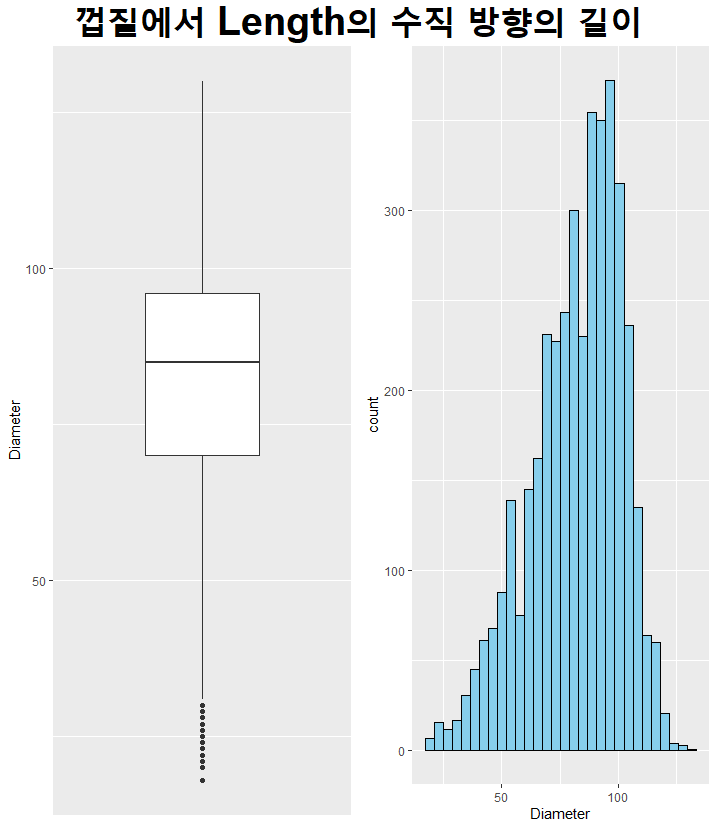
**Ⅳ. 부록**



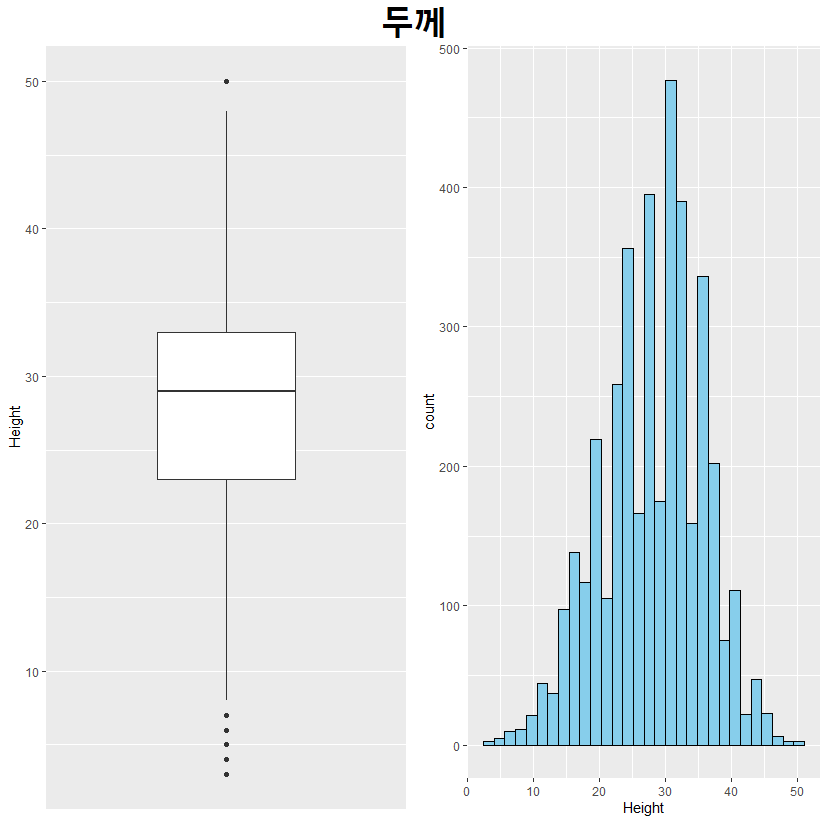
[ 그림 2 – 3 – 1 - 1 ] 로지스틱 회귀 모형의 ROC 곡선



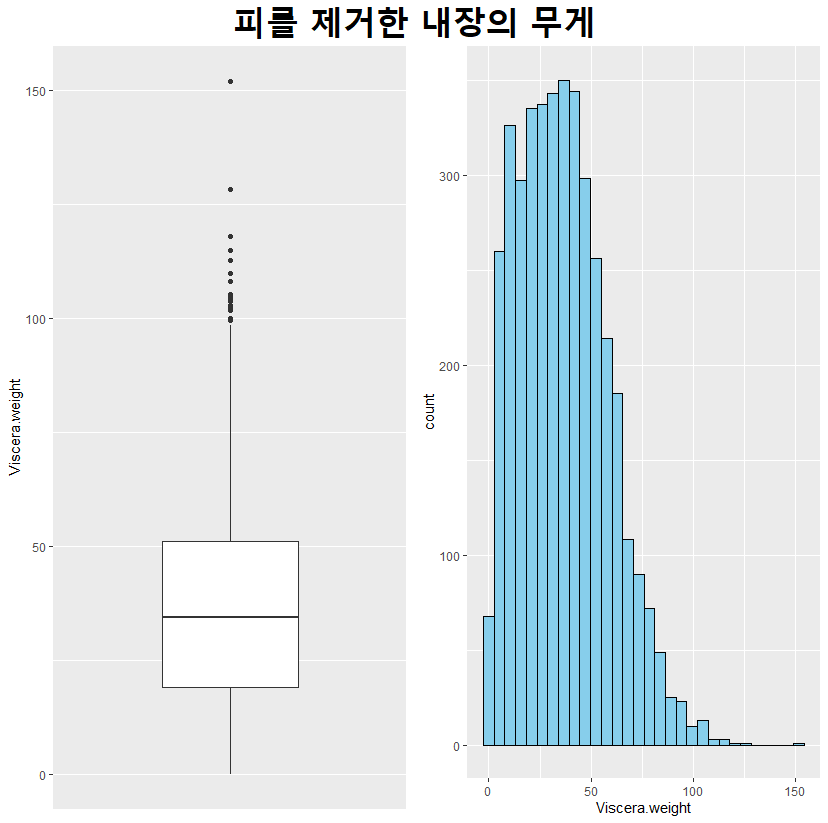
껍질에서 긴 부분의 길이(Length)의 분포



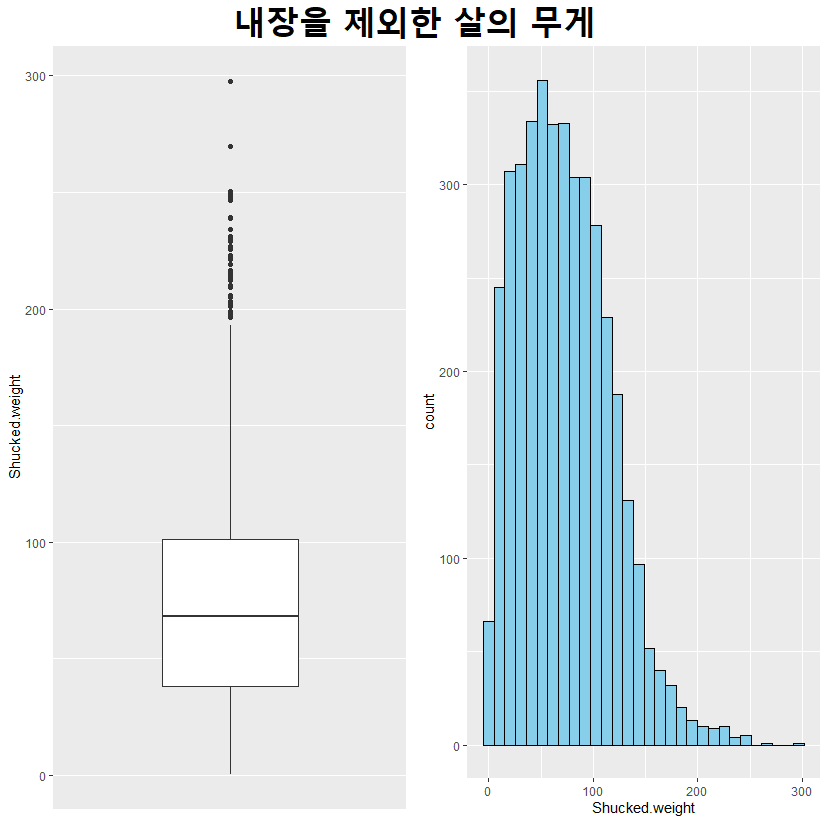
수직방향 길이(Diameter)의 분포



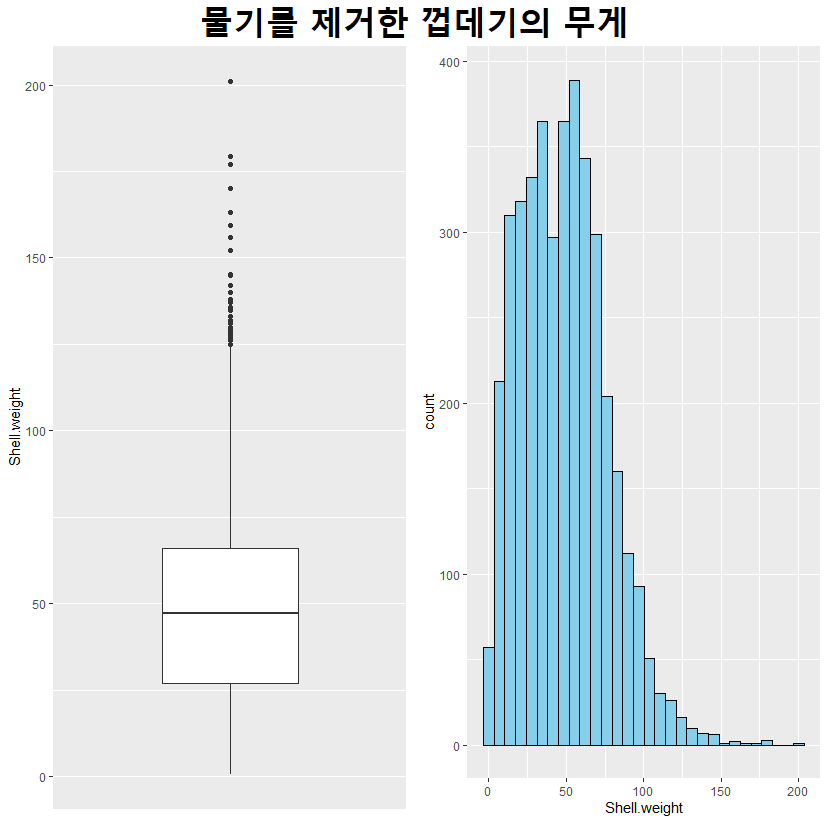
두께(Height)의 분포



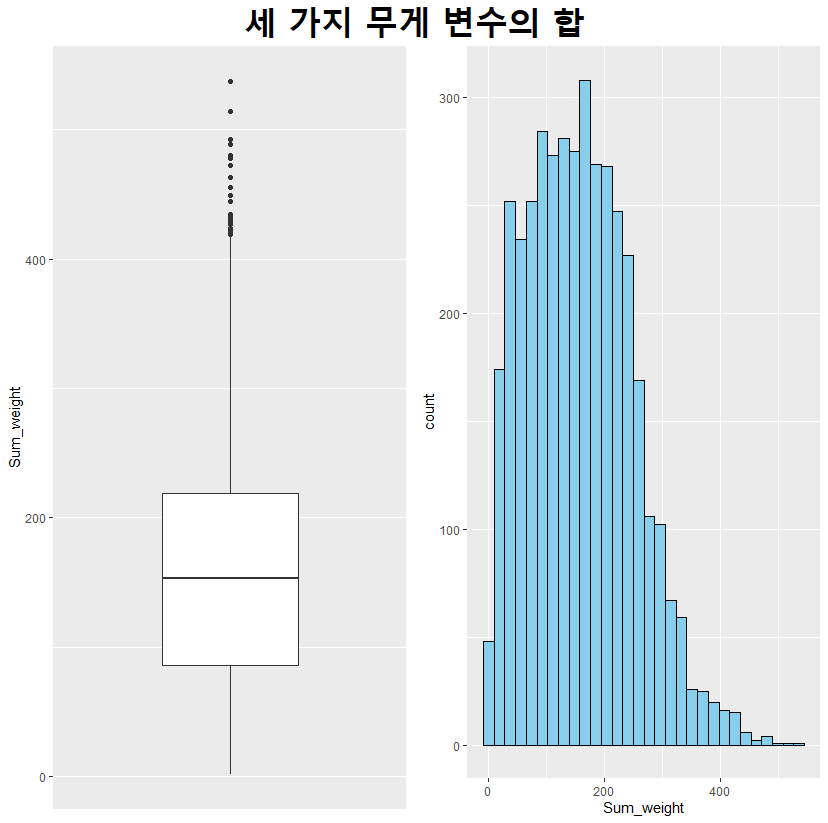
피를 제거한 내장의 무게(Viscera weight)의 분포



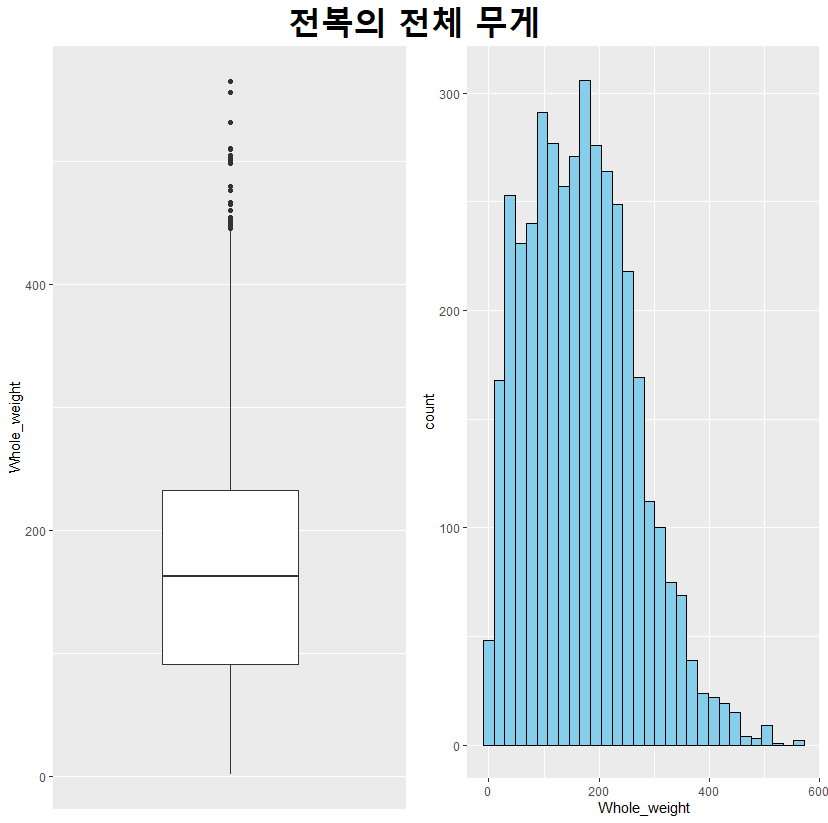
내장을 제외한 살의 무게(Shucked weight)의 분포



물기를 제거한 껍데기의 무게(Shell weight)의 분포



세 가지 무게 변수의 합(Viscera weight+ Shucked weight+ Shell weight)의 분포



전복의 전체 무게의 합(whole weight)의 분포