회귀 기획서

전복 데이터를 사용한 회귀 분석

분석 목적

- 전복의 나이 예측을 통하여 높은 시장 가치를 가진 전복을 평가한다.

대회 개요

- **Regression with an Abalone Dataset**

  Regression with an Abalone Dataset
  Playground Series - Season 4, Episode 4
  https://www.kaggle.com/competitions/playground-series-s4e4/overview

  - 시작일
    - April 1, 2024
  - 마감일 (대회 현재 진행 중)
    - April 30, 2024
  - 평가

    RMSLE(Root Mean Squared Logarithmic Error)
  - 최종 스코어
    - 1st score: 0.14423 (24.4.11 14:24 기준)
  - 목표 스코어
    - Top 20% (Silver)를 목표로 진행한다.

선정된 주제 이슈 조사

▼ 대표적으로 대한민국 완도군의 주력 상품인 전복은 보통 시중에서 팔아서 이윤을 남기기 위해 2년차 성체가 되면 수확을 하며 여기에 전복의 치어라 할 수 있는 치패를 육성하는

데도 반년이 걸리니 판매용이 되려면 2.5년 이상이 걸린다.

- 출처 : 나무위키

https://namu.wiki/w/전복

분석 사례 조사

▼ 전복의 경제적 가치는 연령과 양의 상관 관계가 있다.

https://mpra.ub.uni-muenchen.de/91210/1/MPRA_paper_91210.pdf

출처: MPRA(19.01)

문제 확인

- 전복 나이에 가장 큰 영향을 미치는 변수는 무엇일까?
- 어떤 변수를 선택하고 제거해야할까?
- 변수 전처리는 어떻게 진행해야할까?

데이터

- **Train**

train.csv

  - 행(row) : 90615
  - 열(column) : 10

- 데이터 상위값

	id	Sex	Length	Diameter	Height	Whole weight	Whole weight.1	Whole weight.2	Shell weight	Rings
0	0	F	0.550	0.430	0.150	0.7715	0.3285	0.1465	0.2400	11
1	1	F	0.630	0.490	0.145	1.1300	0.4580	0.2765	0.3200	11
2	2	I	0.160	0.110	0.025	0.0210	0.0055	0.0030	0.0050	6
3	3	M	0.595	0.475	0.150	0.9145	0.3755	0.2055	0.2500	10
4	4	I	0.555	0.425	0.130	0.7820	0.3695	0.1600	0.1975	9

- **Test**

  test.csv

  - 행(row) : 60411
  - 열(column) : 9
  - 데이터 상위값

	id	Sex	Length	Diameter	Height	Whole weight	Whole weight.1	Whole weight.2	Shell weight
0	90615	M	0.645	0.475	0.155	1.2380	0.6185	0.3125	0.3005
1	90616	M	0.580	0.460	0.160	0.9830	0.4785	0.2195	0.2750
2	90617	M	0.560	0.420	0.140	0.8395	0.3525	0.1845	0.2405
3	90618	M	0.570	0.490	0.145	0.8740	0.3525	0.1865	0.2350
4	90619	I	0.415	0.325	0.110	0.3580	0.1575	0.0670	0.1050

- **Sample_submission**
  - 행(row) : 60411
  - 열(column) : 9
  - 데이터 상위값

	id	Rings
0	90615	10
1	90616	10
2	90617	10
3	90618	10
4	90619	10

- Original
  - 행(row) : 4177
  - 열(columns) : 10
  - 데이터 상위값

	id	Sex	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings
0	0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
1	1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
2	2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
3	3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
4	4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7

변수 설명

- id
- Sex: 성별
- Length: 길이
- Diameter: 직경
- Height: 높이
- Whole weight: 전체 무게(살+껍질)
- Whole weight.1: 껍질 벗긴 무게
- Whole weight.2: 내장 중량
- Shell weight: 껍질 무게

문제 해결 전략

- 데이터 전처리
  - 상관관계 분석을 통한 변수 선택
  - 불필요한 피처 제거
  - 인코딩
    - Sex(성별)
      - OneHot Encoder
      - Label Encoder
      - ...
  - 스케일링
    - Standard Scaler
    - Robust Scaler
    - MinMax Scaler
    - ...
- 모델링
  - LightGBM Regressor
  - XGBoost Regressor
  - Lasso
  - ...
- 하이퍼 파라미터 최적화
  - 베이지안 최적화
  - 랜덤서치

일정표

타임라인 (1)

Aa Name	Date	Tags
✔️ 주제 선정	@2024년 4월 11일	
💁‍♀️ 발표	@2024년 4월 26일	
💻 기획서 제출	@2024년 4월 16일	
💻 결과물 제출	@2024년 4월 26일	
✔️ 발표자료 최종 점검	@2024년 4월 25일	
📝 분류/회귀 분석 동시 진행	@2024년 4월 11일 → 2024년 4월 12일	
👩 멘토링	@2024년 4월 20일 오후 7:00	
📝 추가 대회 진행	@2024년 4월 22일 → 2024년 4월 24일	
👩 멘토링	@2024년 4월 22일 오후 8:00	
💻 보고서 작성(피드백 반영)	@2024년 4월 22일	
📝 분류/회귀 분석 동시 진행	@2024년 4월 15일 → 2024년 4월 19일	
💻 보고서 작성	@2024년 4월 19일	

팀원 소개 및 역할

- 김수지 [PM]
  - 기획서 작성
  - 보고서 작성
  - 발표 자료 작성
  - 전복 데이터 세트를 사용한 회귀 분석
    - 데이터 전처리
    - 분석 및 EDA
- 박지은 [PL]
  - 기획서 작성
  - 보고서 작성
  - 발표(회귀)

    - 전복 데이터 세트를 사용한 회귀 분석
        - 모델링
        - 모델 평가
- 변진영 [PM]
    - 기획서 작성
    - 보고서 작성
    - 발표 자료 작성
    - 은행 이탈 데이터를 사용한 이진 분류 분석
        - 모델링
        - 모델 평가
    - 전복 데이터 세트를 사용한 회귀 분석
        - 모델 평가
- 이소희 [PM]
    - 기획서 작성
    - 보고서 작성
    - 발표 자료 작성
    - 은행 이탈 데이터를 사용한 이진 분류 분석
        - 분석 및 EDA
- 이정수 [PL]
    - 기획서 작성
    - 보고서 작성
    - 발표(분류)
    - 은행 이탈 데이터를 사용한 이진 분류 분석
        - 데이터 전처리
        - 모델 평가