2024-11-25 수상작 리뷰

Iris 너비 예측 AI 해커톤

https://dacon.io/competitions/official/235836/codeshare/3705

데이터

1. iris_train.csv : 학습 데이터

- id: 데이터 고유 id
- species: 붓꽃의 종류 (versicolor, setosa, virginica 중 하나)
- sepal length (cm) : 붓꽃의 꽃받침의 길이
- petal length (cm) : 붓꽃의 꽃잎의 길이
- sepal width (cm) : 붓꽃의 꽃받침의 너비
- petal width (cm) : 붓꽃의 꽃잎의 너비

2. test.csv : 테스트 데이터

- id: 데이터 고유 id
- species: 붓꽃의 종류 (versicolor, setosa, virginica 중 하나)
- sepal length (cm) : 붓꽃의 꽃받침의 길이
- petal length (cm) : 붓꽃의 꽃잎의 길이

3. sample_submissoin.csv : 제출 양식

- id: 데이터 고유 id
- sepal width (cm) : 붓꽃의 꽃받침의 너비
- petal width (cm) : 붓꽃의 꽃잎의 너비

코드 흐름

1) **EDA**
   꽃들의 차이점 알기
   패키지 설치
   데이터 확인
   평균 꽃잎과 꽃받침 길이 확인
   붓꽃 종류별로 count
   종류별 꽃받침/ 꽃잎 평균 길이 확인
   시각화를 통해 데이터 살펴보기
   Corr() 확인하기
   ID drop
2) **Feature Engineering & Processing**
   종류 별로 데이터 분리
3) **모델 학습(Linear Regression)**

선형 관계가 존재한다면 LinearRegression()을 사용한다.
선형 관계가 존재하지 않는다면 RandomForestRegressor(max_depth = 1)를 사용한다.
모델을 만들고 실제로 적용해 보았을 때, 예측의 수행이 어느 정도 정확한지 loocv를 통해 관찰한다. 그리고 모델들을 앙상블 한다. 만약 데이터에 이상치가 존재한다면 중앙값을 사용하고, 존재하지 않는다면 평균을 사용한다.

4) 후처리

예측값 할당

차별점, 배울점

데이터의 특성을 잘 파악하기 위해 노력했다. 데이터의 수가 적다는 것은 K-fold로 학습을 수행했을 때, K를 너무 작게 주면 데이터 수의 부족으로 인해서 적절한 모델이 만들어질 수 없음을 제대로 인지하였다. 따라서 loocv(Leave-One-Out Cross-Validation)를 사용해서 모델의 데이터를 최대한 유지해야겠다는 해결책을 내놓았다. 또한, 꽃의 종류에 다라 구분되는 특징이 있음을 인식하고 꽃의 종류에 따른 각각의 학습 모델을 만들었다는 점이 차별점이다.

또한, Loocv란 교차 검증 방식 중 하나로 K-Fold Cross Validation과 다르게 높은 신뢰도를 제공하지만 연산하는 데 시간이 오래걸려서 데이터셋이 매우 작을 때 적합하다.