241129 수상작리뷰

[천체 유형 분류 대회]

☑️**참고 코드**: 천체 유형 분류 대회 1위 수상자 코드 설명 | PPT
https://www.slideshare.net/slideshow/1-231159616/231159616#8

📗 대회 & 데이터 소개

a. 대회 소개:

- 우주는 찰나의 순간에도 천문학적인 양의 데이터를 생산해왔고 그 방대함에 비례하는 데이터를 천문학자들은 수집 및 분석해옴. 슬론 디지털 천체 관측(SDSS)는 우주의 천문학적인 규모의 데이터를 수집함. 데이터 처리에는 머신러닝과 딥러닝 기법이 활용되기 시작.

b. 평가 지표:

- Log Loss

c. 데이터:

- Train 크기: (199991, 24), Test 크기: (10009,22)
- **Target: 19개 천체 Type (다중 Class 분류 문제)**
- psfMag(Point spread function magnitudes): 먼 천체를 한 점으로 가정하여 측정한 빛의 밝기
- fiberMag: 은하처럼 뚜렷한 표면이 없는 천체에서는 빛의 밝기를 측정하기 어려움. 천체의 위치와 거리에 상관없이 빛의 밝기를 비교하기 위한 수치
- modelMag: 천체 중심으로부터 특정 거리의 밝기
- fiberID: 관측에 사용된 광섬유의 구분자

📗 코드흐름

1. **데이터 전처리 & EDA**
   - (타입 별 hist)일부 타겟 Class는 매우 드물다.
   - (psfMag_u, psfMag_g 각각의 hist) 타겟 별 매우 다른 magnitude분포를 가진 것으로 보아 원본 magnitude 분포도 좋은 피처가 될 것 같다.
   - (2개의 Filter difference를 scatterplot) target들이 구분됨
   - (SDSS 알고리즘) r-I, g-r 같은 diff feature들로 천체를 구분
   - 모든 Magnitude의 같은 filter u를 Target 별로 살펴봤을 때 대부분의 분포가 Magnitude에서 비슷하지만 일부 Target은 구분됨.
   - Legacy Galaxy 알고리즘에서도 다른 Magnitude filter끼리 빼는 것을 볼 수 있음
   - Magnitude의 첨도가 다르므로 표준편차로 타겟 구분 가능
   - 일부 타겟은 중심값의 위치가 다름

- Max Magnitud, Min Magnitude의 filter가 다름, 차이도 큼. filter별로 겹치지 않는 것도 있음
- 다양한 피처 추가:
    - Magnitude, row 별 max-min, std, sum Feature 추가
    - 모든 magnitude들의 조합으로 diff 피처 추가
    - 각 magnitude별 max-max, min-min, sum-sum 추가
    - Ugriz -> sdssUBVRITransform
    - fiberID 별 fiberMag mean, (fiber_u, g,r, i, z) / fiberMag_mean
    - SDSS 문서를 바탕으로 피처 추가
    - Asinh 변환
    - Diff 피처들의 표준 편차
    - Original Magnitude만 차원 축소하여 피처로 사용

2. 모델 구축 & 검증: LGBM DART Single Model

- Permutation Importance: 피처의 영향도 측정, random-noise 추가**
- LGBM 파라미터는 HyperOpt로 찾음
- 최종 LGBM DART single model
    - DART: Gradient Boosting 알고리즘은 나중에 추가된 트리는 큰 기여를 하지 못하고 성능에 부정적인 영향을 미치는 over-specialization이 있는데 DART는 dropout을 사용해 해결함. Early stopping을 사용할 수 없음. 속도가 느리므로 보통 gbdt로 피처 엔지니어링을 하고 ㄴ앙상블 시 gbdt+dart 사용
- StartifiedKFold 사용(5fold 검증): 각 폴드마다 타겟의 비율이 비슷하게 분배되게 함

3. 제출

📗 배운점 / 특이점

- 다중 class 분류 문제를 처리하는 흐름을 알 수 있었다.
- 모델링보다 피처엔지니어링이 중요한 부분이었다한만큼 도메인 지식을 알수록 도움이 되는 것 같다
- 분포를 시각화해 타겟 타입이 구분되는 피처를 중요 피처로 파악한다.
- 관련 문서를 참고해 다양한 피처를 추가해본다
- **Permutation Importance: score가 얼마나 감소하는지 측정해 피처 영향도를 측정. 전체 데이터로 먼저 학습하고 예측 시 원본 피처를 제거한 피처 값에 랜덤 노이즈를 추가해 학습하면 피처의 기능을 잃으므로 이때의 점수 감소 정도를 확인하는 방법임(cutoff 값도 구함)
- GBDT, GOSS, DART 모델