ISIC 2024 - Skin Cancer Detection with 3D-TBP

ISIC 2024 수상작 리뷰

발표 순서

01 주제 선정 이유
아이데이션

02 캐글 대회 소개
대회 TASK, Dataset 소개

03 솔루션 리뷰
리더보드 상위 5개 솔루션 리뷰

04 시도한 모델링
캐글 submission

05 프로젝트 성과
프로젝트 성과 공유

06 느낀점
팀원 소감 및 마무리

주제 선정 이유

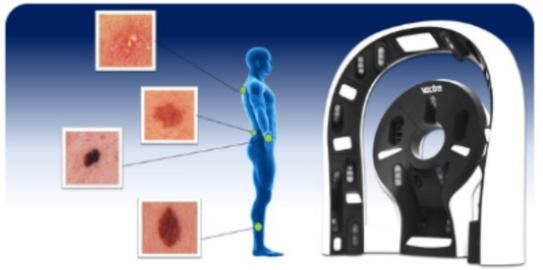
Kaggle 수상작 리뷰를 통한 공모전 참여 경험 쌓기 & 수상 노하우 탐구

메타데이터 이해(jpeg 형식의 피부암 이미지 데이터 등) 및 이미지 분류 모델링 경험 쌓기

캐글 대회 소개

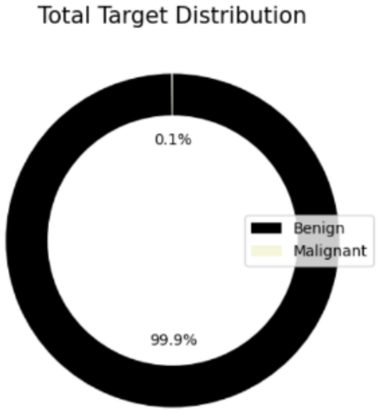
ISIC 2024 - Skin Cancer Detection with 3D-TBP

Identify cancers among skin lesions cropped from 3D total body photographs



: 피부병변 중 악성/양성 데이터를 분류하는 대회

- 이미지 데이터와 Tabular 데이터를 모두 사용 필요
- Target 데이터가 매우 불균형한 문제 해결 필요
- 주어진 데이터의 피처 수가 매우 많아 피처 엔지니어링 필요



대회 데이터 소개

+

	isic_id	target	patient_id	age_approx	sex	anatom_site_general
0	ISIC_0015670	0	IP_1235828	60.0	male	lower extremity
1	ISIC_0015845	0	IP_8170065	60.0	male	head/neck
2	ISIC_0015864	0	IP_6724798	60.0	male	posterior torso
3	ISIC_0015902	0	IP_4111386	65.0	male	anterior torso

병변에 대한 Image Data + 해당 병변(+환자)에 대한 Tabular Data

Tabular Data

anatom_site_general	Location of the lesion on the patient's body. 병변이 위치한 신체 부위.
clin_size_long_diam_mm	Maximum diameter of the lesion (mm). + 병변의 최대 직경.
image_type	Structured field of the ISIC Archive for image type. 이미지 유형(구조화 된 Field).
tbp_tile_type	Lighting modality of the 3D TBP source image. 3D TBP 원본 Image의 조명 방식
tbp_lv_A	A inside lesion. + 병변 내부의 A 값.
tbp_lv_Aex	A outside lesion. + 병변 외부의 A 값.
tbp_lv_B	B inside lesion. + 병변 내부의 B 값.
tbp_lv_Bext	B outside lesion.+ 병변 외부의 B 값.
tbp_lv_C	Chroma inside lesion.+ 병변 내부의 색도.
tbp_lv_Cext	Chroma outside lesion.+ 병변 외부의 색도.
tbp_lv_H	Hue inside the lesion, calculated as the angle of A *and B* in LAB* color space. T to 75 (brown). + 병변 내부의 색상.
tbp_lv_Hext	Hue outside lesion. + 병변 외부의 색상.
tbp_lv_L	L inside lesion. + 병변 내부의 명도.
tbp_lv_Lext	L outside lesion. + 병변 외부의 명도.
tbp_lv_areaMM2	병변의 면적.

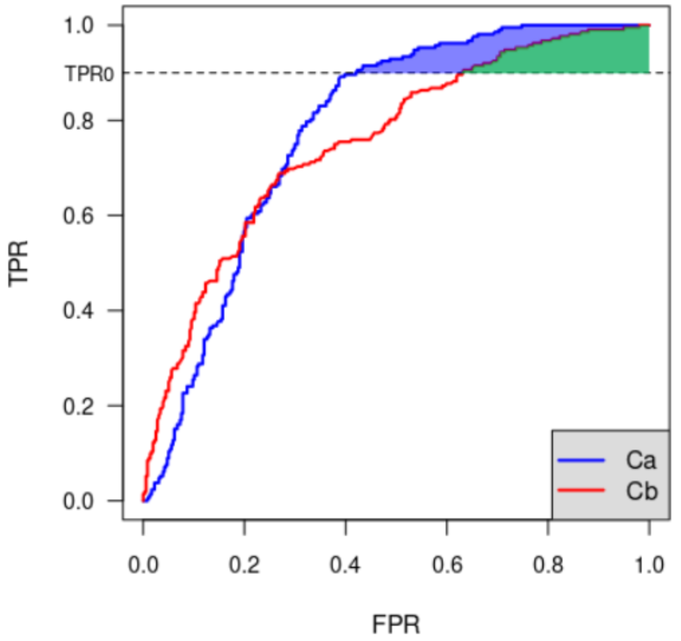
- 칼럼의 수가 매우 많아 피처 엔지니어링 필요
- 이미지에 대한 수치적 정보 다수 포함
  - ⇒ (이미지 + 태블러) VS (Only 태블러)의 분석 결과 비교 필요성 인식

다수의 칼럼이 병변 이미지에 대한 수치

평가지표 - pAUC

**:ROC 곡선의 일부만을 평가하는 지표로,
특정 구간에서의 모델 성능을 집중적으로 분석하는 방법**

1. ROC 곡선 계산
2. TPR이 80% 이상인 구간만 선택
3. 해당 구간의 AUC(면적) 계산



사용 목적

- 의료 분야에서 중요한 민감도(TPR 80% 이상)를 반영
- 민감도가 낮은 모델을 벌점 부여

Top 5 솔루션 리뷰

공통

- ✓ CatBoost, LGBM, XGB 모델링을 합쳐둔 GBDT 모델을 사용해 예측
- ✓ Optuna tuning으로 fine tuning 진행
- ✓ 여러 부스팅 모델을 앙상블해서 최종 결과값 예측
- ✓ 전반적으로 작은 이미지 모델(eva02_small) 사용
- ✓ 이미지 모델로 예측 수행 후 해당 예측 값을 GBDT 모델의 Feature로 사용

Top 5 솔루션 리뷰

개별

- ✓ Diffusion model로 synthetic dataset 생성
- ✓ Mixup augmentation
- ✓ 이전 대회 데이터를 사용해 데이터 보충

} 불균형 데이터 처리 방식

- ✓ Tabular ugly duckling
- ✓ 더 세부적으로 분류된 데이터 라벨 활용

} 기타 성능 향상 테크닉

모델링 결과 정리

Image	Model	Private	Public
Image O	LGBM + ImageNet	0.16528	0.18384
	CatBoost + ImageNet	0.16138	0.18019
	XGBoost + ImageNet	0.16561	0.18043
Image X	LGBM + CatBoost	0.15048	0.17739
	LGBM	0.16579	0.18135

✔ 이미지 데이터를 사용하지 않고도 높은 점수

∵ Tabular 데이터에는 피부 병변의 중요한 수치적 정보(병변 크기, 명도 등)가 이미 포함

프로젝트 성과

1. GitHub 및 코드 관리
    - config, src 폴더 구조를 이해하고 코드 정리 및 재사용성을 높이는 방법 학습
    - .ipynb 중심의 작업에서 .py 파일 중심의 코드 작성 및 분석 방식으로 확장
    - Kaggle 대회에서 GitHub을 활용한 협업 방식 및 코드 공유 방법 습득
2. Kaggle 커뮤니티 및 정보 공유
    - Discussion 탭을 활용해 다양한 솔루션을 참고하고, 이를 바탕으로 최적의 모델을 구현하는 경험
    - Kaggle 대회에서 가장 효율적인 정보 공유 및 협업 방식 이해
    - GitHub에 기록된 프로젝트 과정과 하이퍼파라미터 튜닝 결과물(yaml 파일)을 분석하는 방법 습득
3. Submission & Private Score 비교 분석
    - Kaggle 대회에서 모델 제출 후 Public/Private Score 비교 및 디버깅 방법 이해
    - pAUC(Partial AUC)라는 새로운 평가 지표를 학습하고 TPR(민감도) 기반의 모델 평가 전략 습득
    - 최적의 모델을 찾기 위해 하이퍼파라미터 튜닝 및 앙상블 기법 활용

프로젝트 성과

4. 데이터 불균형 문제 해결 및 Loss 조정
    - Good Under-sampling, Over-sampling 같은 데이터 샘플링 기법 학습
    - Loss 함수 조정 및 Cost-sensitive learning을 통해 불균형 데이터에서도 모델이 안정적으로 학습되도록 조절
    - 다양한 모델링 기법을 실험하며 데이터 불균형 문제를 해결하는 최적의 전략 탐색
5. 데이터 증강(Augmentation) 및 최적화 기법 학습
    - Augmentation 방법론 및 최적의 Feature Engineering 전략 실험
    - 모델 성능을 극대화하기 위한 다양한 전처리 및 후처리 기법 적용
6. 다양한 수상작 코드 리뷰 및 적용
    - Kaggle 상위권 솔루션을 분석하면서 새로운 테크닉과 모델링 전략을 적용해봄
    - .ipynb이 아닌 .py 파일을 중심으로 코드를 리뷰하면서 더 체계적인 코드 분석 및 적용 방법을 배움
    - 다양한 모델 조합을 실험하며 LGBM, CatBoost, XGBoost 등 다양한 기법을 비교 분석
7. Tabular 데이터만으로도 강력한 모델 구축 가능성 확인
    - ImageNet을 사용하지 않고도 Tabular 데이터만으로 높은 성능을 기록
    - Tabular 데이터에 크기, 모양, 색상, 이심률 등의 정보를 반영해 이미지 없이도 성능이 우수한 모델 개발 가능

프로젝트 소감

단순한 모델링만 진행하는 것이 아닌, 더 세부적으로 데이터를 살펴보고 처리하는 과정이 분석에 필요하다는 것을 느낌.
그 외에 데이터 불균형을 다루는 방법들에 대해 더 배울 수 있는 프로젝트였음

kaggle 대회 참가자들이 서로 정보를 공유하는 discussion에 대해 알게 되었고, github에 기록된 프로젝트 과정과 코드, 튜닝 결과물(yaml 파일) 등을 파악하는 방법을 터득할 수 있는 유익한 프로젝트

수상작 코드 리뷰 시 깃허브에서 .py 파일을 다루는 것을 처음 진행해봄. 데이터 불균형 해결 방법에 대해 다양하게 생각해볼 수 있었음. pAUC라는 새로운 평가 지표를 알게 되는 등 많은 것을 배운 프로젝트

수상작 리뷰를 통해서 다양한 코드를 이해하는 능력을 더 키울 수 있었고, 모델링 과정에서 모델을 바꾸는 것뿐만 아닌 OOF 분석 방법 등의 성능 최적화 방법을 배울 수 있던 프로젝트

이상으로 발표를 마칩니다.