유방암의 임파선 전이 예측
AI 경진대회

aivle team 김시연, 이동비, 조진호, 엄유정

목차

1. 분석결과 요약
2. 데이터 전처리
3. 모델링
4. 시도사항
5. 정확도

1. 분석결과 요약

분석결과 정확도

- XGBoost : 0.80766212 (public)
- XGBoost + CNN : 0.8235578125 (public)

분석 흐름도

- 각 데이터에 대한 분석, 전처리 과정후 XGBoost, CNN 모델링
- 이미지와 결측치에 대한 다양한 처리 방법 시도

2. 데이터 전처리 - 임상 항목(정형) 데이터 결측치 처리

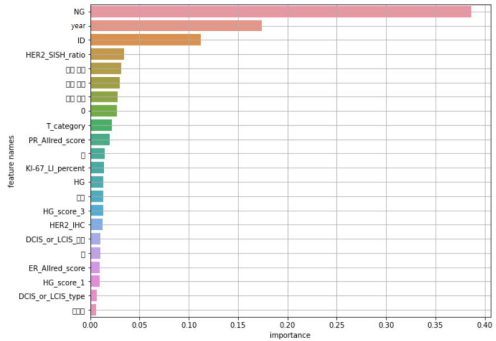
중요 변수 추출 후
KNN IMPUTER 적용

암의 장경
NG
HG
ER_Allred_score
PR_Allred_score
KI-67_LI_percent
HER2_SISH_ratio

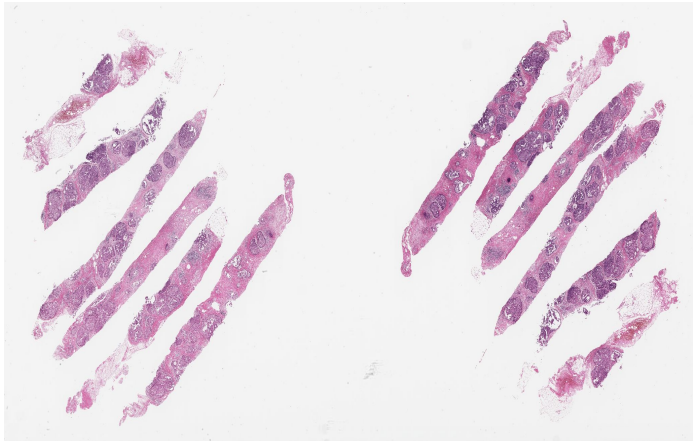
수술연월일 ->
년,월,일 분리

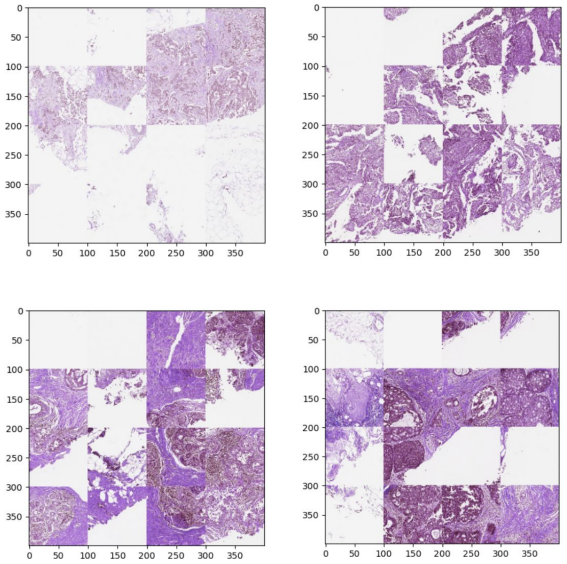
중요도 낮은
컬럼을 제거

남은 변수
0으로 채우기



2. 데이터 전처리 - 유방암 병리 슬라이드 영상 이미지

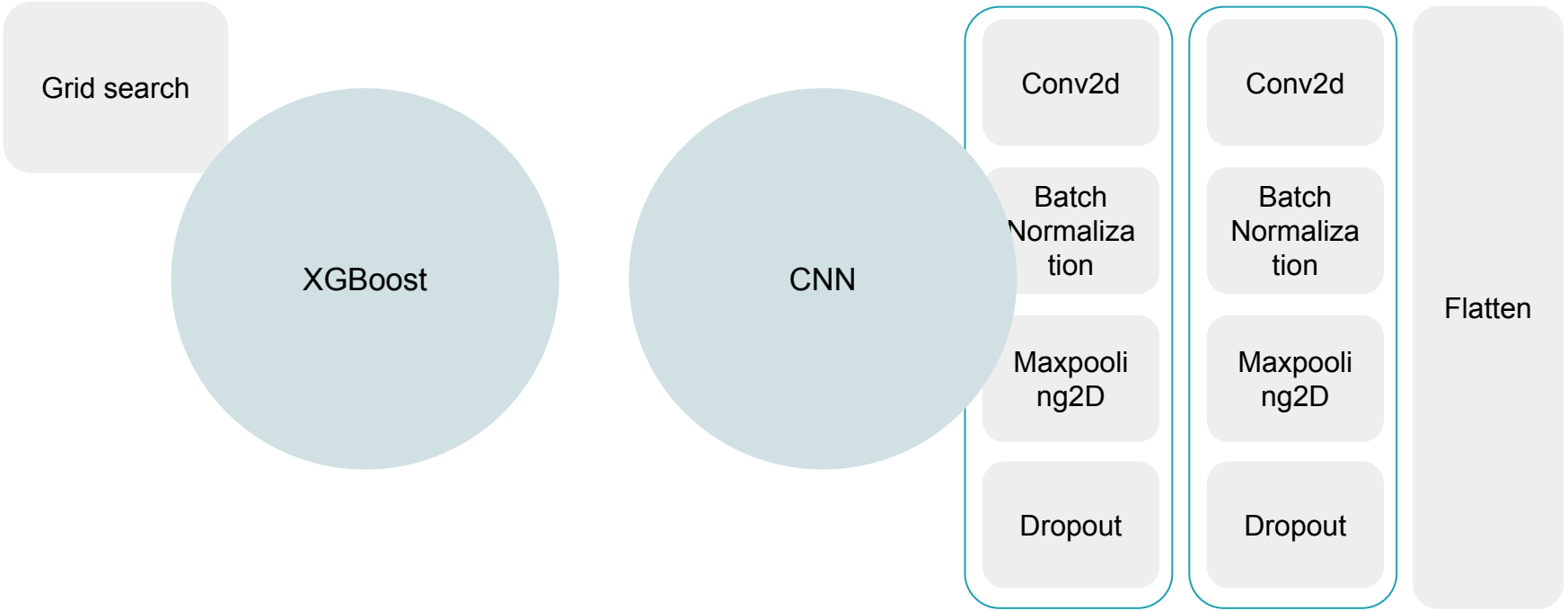




-여러 실험과정을 통해 가장 정확도가 높은 이미지 추출방식 적용-

1. 이미지를 grayscale 변환
2. 0,0위치부터 오른쪽으로 차례로 탐색
3. 픽셀 값이 200이 넘는 지점이 나오면 , 그 지점으로부터 100x100 크기로 이미지 캡처
4. 캡처 후 오른쪽, 아래로 100 이동후 3 반복
5. 캡처된 이미지가 16장이 되었을 경우 400x400이미지로 결합

3. 모델링


Grid search
XGBoost
CNN
Conv2d
Batch Normaliza tion
Maxpooli ng2D
Dropout
Conv2d
Batch Normaliza tion
Maxpooli ng2D
Dropout
Flatten

4. 시도사항

XGBoost

문제점 : 데이터가 낯설어서 처음 데이터를 분석할 때 전처리를 접근 방법이 어려웠음

시도 1 : 결측치 컬럼 전부 0으로 채우기 -> 0.779

시도 2 : 중요도 낮은 컬럼 제거, 남은 결측치 변수들 0으로 채우기 -> 0.807

시도 3 : 중요도 높은 컬럼에 KNN IMPUTER 적용, 중요도 낮은 컬럼 제거, 남은 결측치 변수들 0으로 채우기 -> 0.829

CNN

문제점: 학습 이미지 크기가 매우 크고 달라서 학습환경에서 메모리 부족 문제가 발생

시도1:이미지를 reshape시켜서 400x400크기로 조정
=> 학습 정확도가 0.5 수준으로 학습 불가능

시도2:이미지의 세포부분만 400x400크기로 캡처해서 학습
=>전체 이미지중 아주 일부분만 가져와 학습하게 되어 학습 정확도가 0.53수준으로 매우 떨어짐

시도3:전체이미지를 골고루 100x100사이즈로 16장 캡처하여 이어붙인 후 학습
=> 정확도가 0.7으로 높은 정확도를 가져 **해당 방식을 채택**
캡처 이미지수를 더 많이 할수록 높은 정확도 예상되나 메모리 문제로 시도 하지 못하였음

5. 정확도

XGBoost 모델 : 0.80766212 (public score 기준)

CNN 변수 추가 XGBoost 모델 : 0.8235578125 (public score 기준)

ID	img_path	mask_path	나이	수술연월일	진단명	암의 위치	암의 개수	암의 장경	NG	HG	HG_score_	HG_score_	HG_score_	DCIS_or_LI	DCIS_or_LI	T_categori	ER	ER_Allred_	PR	PR_Allred_	KI-67_LI_p	HER2	HER2_IHC	HER2_SISH	HER2_SISH	BRCA_mu	N_categc	CNN
BC_01_00	./train_img	-	63	2015-10-23	1	2	1	19	2	1	2	2	1	2		1	1	8	1	6	12	0	1				0	2.38E-32
BC_01_00	./train_img	-	51	2015-10-28	1	1	1	22	3	3	3	3	3	0		2	0		0		70	0	0					0.72539
BC_01_00	./train_img	-	37	2015-10-29	1	2	1		2					1	2	0	1	7	1	4	7	0	1			0	0	0.68973
BC_01_00	./train_img	-	54	2016-03-08	1	2	1	0	3	3	3	3	2	1	2	0	0		0		1	1	3				0	0.5759
BC_01_00	./train_img	-	57	2015-10-30	1	2	1	8	2	2	3	2	1	2		1	1	8	0		8	1	2	1	5.44		0	0.59463

CNN 결과
변수로 추가

감사합니다