2021-2

자료분석특론 기말 프로젝트

212STG18 예지혜

Part A

Levels Fyi Salary data

Part A

데이터 소개

Levels fyi Salary data

데이터 일부

timestamp	company	level	title	totalyearlycompensation	location	yearsofexperience
6/7/2017 11:33:27	Oracle	L3	Product Manager	127000	Redwood City, CA	1.5
6/10/2017 17:11:29	eBay	SE 2	Software Engineer	100000	San Francisco, CA	5.0
6/11/2017 14:53:57	Amazon	L7	Product Manager	310000	Seattle, WA	8.0
6/17/2017 0:23:14	Apple	M1	Software Engineering Manager	372000	Sunnyvale, CA	7.0
6/20/2017 10:58:51	Microsoft	60	Software Engineer	157000	Mountain View, CA	5.0

basesalary	stockgrantvalue	Doctorate_Degree	Highschool	Some_College	Race_Asian	Race_White
107000.0	20000.0	0	0	0	0	0
0.0	0.0	0	0	0	0	0
155000.0	0.0	0	0	0	0	0
157000.0	180000.0	0	0	0	0	0
0.0	0.0	0	0	0	0	0

데이터 설명

Data Science와 STEM 분야의 연봉 데이터
Levels.fyi 사이트에서 스크래핑한 데이터

주요 변수

회사	1631개 종류, 관측치가 적은 회사가 매우 많음
경력, 회사 내 경력	연속형 변수
성별	여성, 남성, 기타 3가지 범주, 30%의 결측치
학력	고등학교 졸업, 박사 학위 등 5가지 범주, 50%의 결측치
인종	아시안, 백인, 흑인 등 5가지 범주, 50%의 결측치
직급	비정제된 경우 많음, 일부 결측치
직무	15가지 범주이나 65%가 소프트웨어 엔지니어

분석 목표

관측치 Top 5 회사의 연간 보상 예측
[아마존, 마이크로소프트, 구글, 페이스북, 애플]

Part A

주요 EDA 및 전처리

Levels fyi Salary data

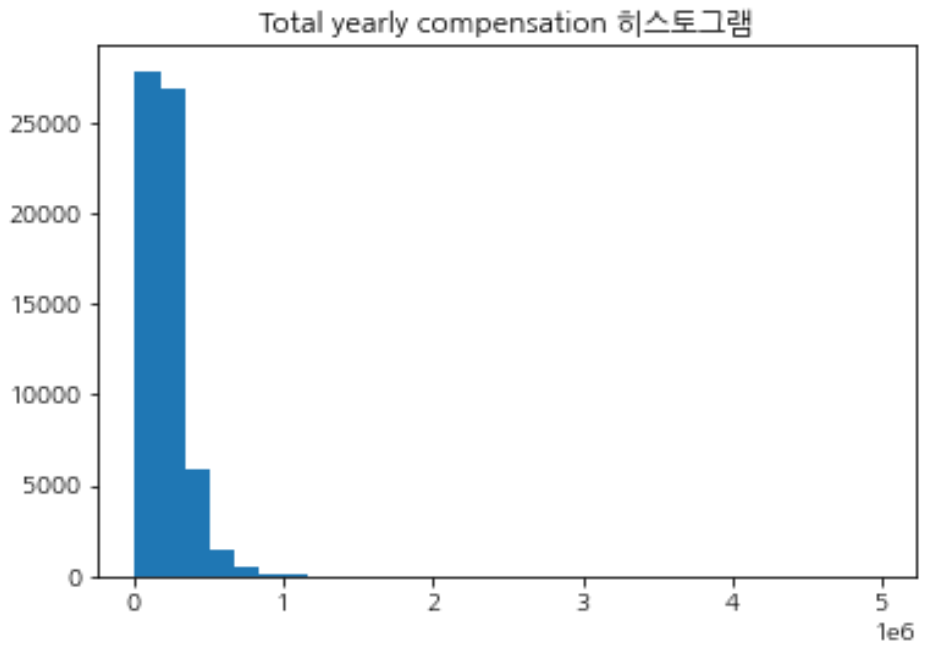
타겟 변수 설정

	연간 전체 보상	기준 연봉	스톡 그랜트	보너스
Mean	216,300	136,687	51,486	1,9335
Std	138,034	61,369	81,874	26,781
Min	10,000	0	0	0
Median	188,000	140,000	25,000	14,000
Max	4,980,000	1,659,870	2,800,000	1,000,000

- 기준 연봉에 0이라는 관측치 존재
- '연간 전체 보상'이 나머지 세 정보의 합산인 경우 80%, 아닌 경우 20%

➔ '연간 전체 보상'을 타겟 변수로 설정

'연간 전체 보상' 히스토그램



Part A

주요 EDA 및 전처리

Levels fyi Salary data

직급 변수 전처리

출처 : Levels fyi

전처리 과정

1. 딕셔너리를 사용해 숫자가 포함된 직급 정보로 수정
2. 숫자 정보만 뽑음
3. 마이크로소프트의 경우, 59부터 2로 간주하여 수정
4. 결측치는 가장 많은 직급인 5단계로 imputation

Part A

주요 EDA 및 전처리

Levels fyi Salary data

전처리

회사	원핫인코딩
성별	결측치는 '기타' 범주로 간주, 원핫인코딩
학력	원핫인코딩된 변수 사용
인종	원핫인코딩된 변수 사용
지역	관측치 1000개 이상인 8개만 살리고, 나머지는 '기타'로 간주, 원핫인코딩

최종 데이터

22,690개의 관측치와 46개의 설명변수
년도, 직급, 경력, 회사내 경력, 학력, 인종, 회사, 지역, 직무, 성별 총 10가지 정보

Part A

모형 탐색 과정

Levels fyi Salary data

9가지 후보 모델

- 선형 회귀, 릿지, 라쏘, 의사결정트리, 랜덤포레스트, 로지스틱 회귀, svm, svr, 그래디언트 부스팅
- 성능이 좋은 모델 그리드 서치

9가지 후보 모델 – 로그 변환

- 선형 회귀, 릿지, 라쏘, 의사결정트리, 랜덤포레스트, 로지스틱 회귀, svm, svr, 그래디언트 부스팅
- 성능이 좋은 모델 그리드 서치

모형 탐색

- 주로 트리 기반 모델의 성능이 좋아 최근 개발된 캣부스트 추가
- 범주형 변수 원핫 인코딩 vs. 라벨 인코딩

추가 모델 - 캣부스트

- 직급 정보 사용 여부
- 학력 정보 원핫 인코딩 vs. 순서 인코딩

그외 고려사항

Part A

모형 탐색 과정 - 9가지 후보 모델

Levels fyi Salary data

모델 종류	파라미터	Test RMSE	Test MAE
선형 모델		101,100	62,688
릿지 회귀	(디폴트) α = 1	101,100	62,681
	α = 10	101,102	62,630
라쏘 회귀	(디폴트, 그리드서치) α = 0.1	101,100	62,687
의사결정트리		93,731	56,005
랜덤포레스트	(디폴트) max_depth = None, min_samples_leaf = 1, n_estimators = 100	75,483	44,168
	max_depth = 50, max_features = 8, min_samples_leaf = 2, n_estimators = 30	76,330	44,178
	max_depth = 50, max_features = 10, min_samples_leaf = 2	76,382	44,007
로지스틱 회귀		135,661	80,428
SVM		128,195	76,066
SVR	Kernel="poly", degree = 2, C=100, epsilon = 0.1, gamma="scale"	148,814	91,829
그래디언트 부스팅	max_depth=2, n_estimators=100, learning_rate=1.0	82,753	48,384
	max_depth=10, n_estimators=50, learning_rate=0.2	74,263	43,042
	max_depth=10, n_estimators=40, learning_rate=0.2	74,216	43,031

Part A

모형 탐색 과정 - 9가지 후보 모델 (로그 변환)

Levels fyi Salary data

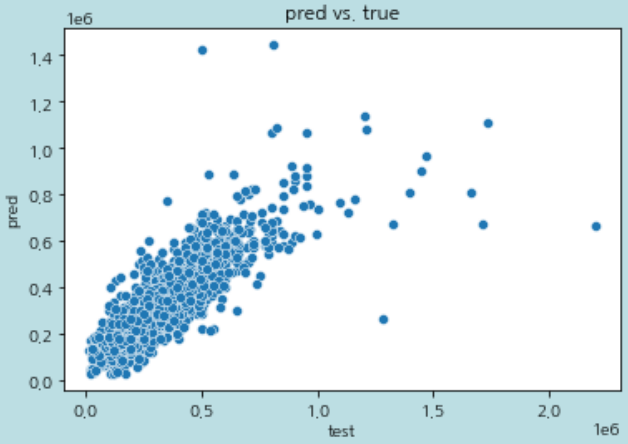
모델 종류	파라미터	Test RMSE	Test MAE
선형 모델		**98,835**	**57,164**
릿지 회귀	(디폴트) α = 1	**98,841**	**57,163**
라쏘 회귀	(디폴트) α = 0.1	129,074	78,119
의사결정트리		93,463	56,530
랜덤포레스트	(디폴트) max_depth = None, min_samples_leaf = 1, n_estimators = 100	76,705	44,308
	max_depth = 20, max_features = 8, min_samples_leaf = 10, n_estimators = 30	85,149	46,947
SVM		**106,841**	**63,584**
SVR	Kernel="poly", degree = 2, C=100, epsilon = 0.1, gamma="scale"	**118,274**	**68,727**
그래디언트 부스팅	max_depth=2, n_estimators=100,learning_rate=1.0	94,604	45,777
	max_depth=5, n_estimators=50, learning_rate=0.5	77,689	43,598

Part A

모형 탐색 과정 – 최종 모델 후보

Levels fyi Salary data

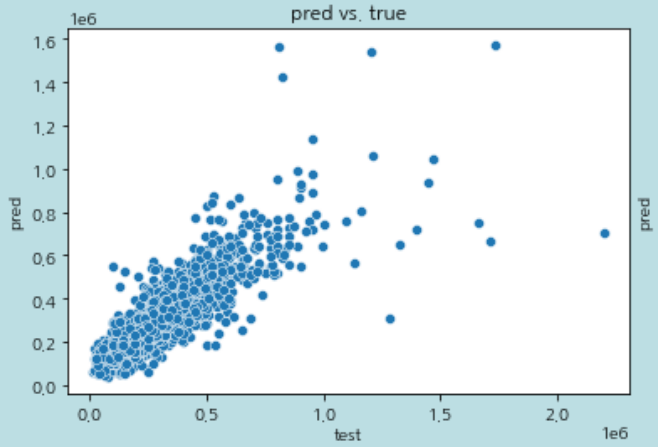
랜덤포레스트



Test RMSE 75,483

Test MAE 44,308

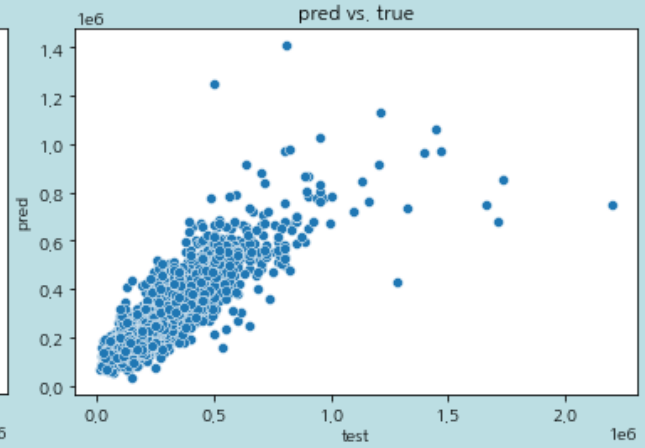
그래디언트 부스팅



Test RMSE 74,216

Test MAE 43,031

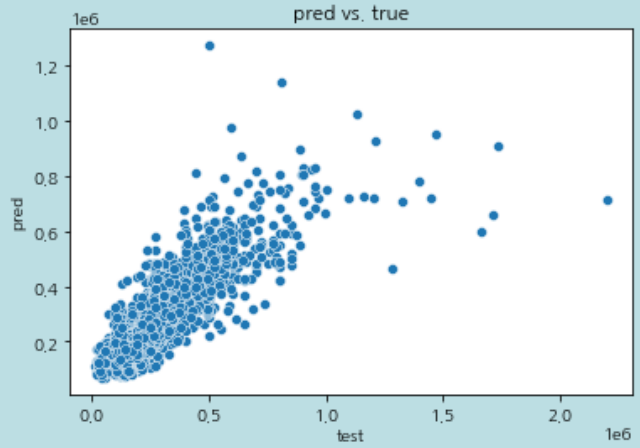
캣부스트 - 원핫인코딩



Test RMSE 73,806

Test MAE 43,553

캣부스트 - 라벨인코딩



Test RMSE 76,960

Test MAE 44,413

Part A

최종 모델

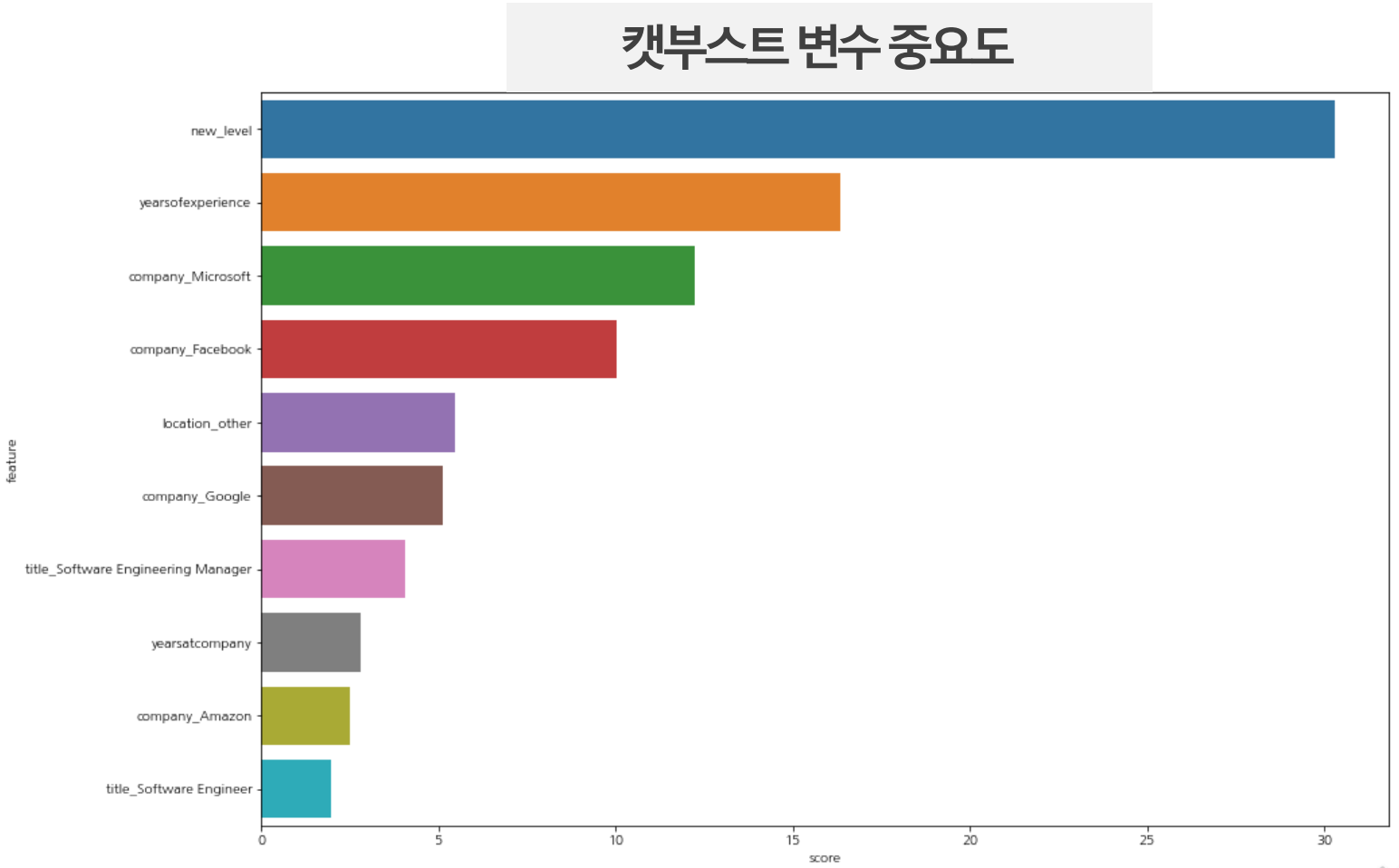
Levels fyi Salary data

모델 설명

- 캣부스트는 범주형 변수를 자체적인 알고리즘에 기반해 효과적으로 처리
- 경력 정보 2개를 제외하고 모두 범주형 변수이기 때문에 부스팅 모형에서 좋은 성능을 보이고, 특히 캣부스트를 사용하기 좋은 데이터임

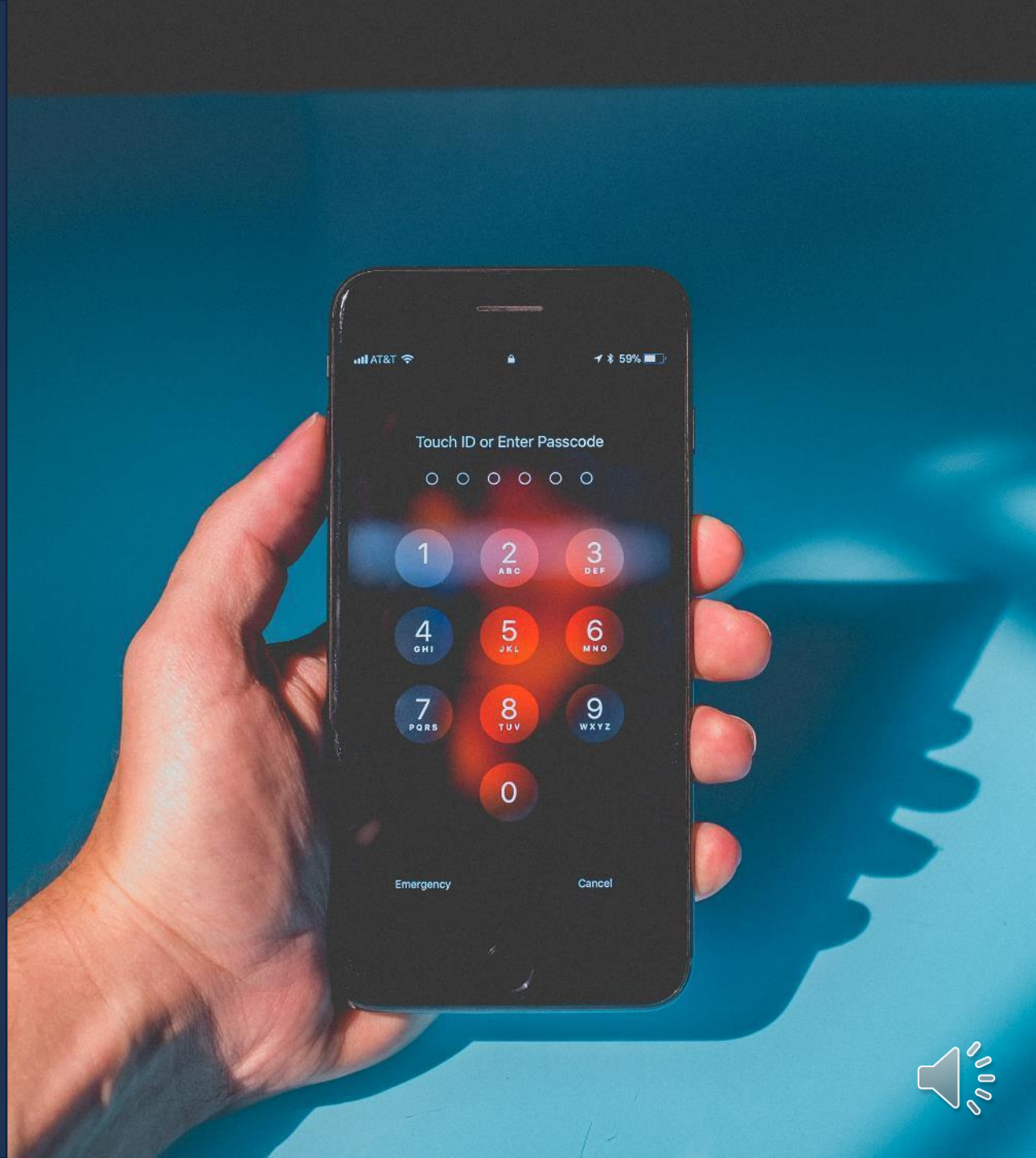
중요 변수

- 전처리된 직급
- 경력
- 회사
- 직무



Part B

Mobilephone Image data



Part B

데이터 소개

Mobile phone Image data

데이터 일부

Model	Brand	Condition	Image_File
iPhone 7	Apple	Used	mobile_images/1635051927882_Apple iPhone 7 Por...
iPhone 6S	Apple	Used	mobile_images/1635051928230_Apple iPhone 6S 12...
Galaxy M02	Samsung	Used	mobile_images/1635051928415_Samsung Galaxy M02...
iPhone 7	Apple	Used	mobile_images/1635051928818_Apple iPhone 7 128...
Galaxy M02	Samsung	New	mobile_images/1635051929034_Samsung Galaxy M02...








데이터 설명

스마트폰 외관 사진 데이터
거래 사이트 Ikman.lk에서 스크래핑한 데이터

데이터 형태

모델	462 종류
브랜드	32 종류

이미지 파일 경로가 path 형태로 csv에 저장

분석 목표

애플, 삼성 두 가지 브랜드로 분류

Part B

전처리

Mobile phone Image data

csv 파일의 이미지 경로 중 실제 이미지 파일이 없는 경우 삭제

애플과 삼성 두 폴더로 이미지 이동

train, val, test 데 이터셋 분리

디렉토리 형태

Data_for_cnn	train	Apple
		Samsung
	val	Apple
		Samsung
	test	Apple
		Samsung

Part B

모델 탐색 과정

Mobile phone Image data

간단한 합성망과 이미지 보강

간단한 합성망 (인풋 150 × 150)		
Layer type	Output Shape	Param #
Conv2D	(None, 148, 148, 32)	896
MaxPooling2D	(None, 74, 74, 32)	0
Conv2D	(None, 72, 72, 64)	18496
MaxPooling2D	(None, 36, 36, 64)	0
Conv2D	(None, 34, 34, 128)	73856
MaxPooling2D	(None, 17, 17, 128)	0
Conv2D	(None, 15, 15, 128)	147584
MaxPooling2D	(None, 7, 7, 128)	0
Flatten	(None, 6272)	0
Dense	(None, 512)	3211776
Dense	(None, 1)	513
Total params: 3,453,121		
Test acc : 0.730 / 이미지 보강시 0.665		

간단한 합성망 (인풋 255 × 255)		
Layer type	Output Shape	Param #
Conv2D	(None, 253, 253, 32)	896
MaxPooling2D	(None, 126, 126, 32)	0
Conv2D	(None, 124, 124, 64)	18496
MaxPooling2D	(None, 62, 62, 64)	0
Conv2D	(None, 60, 60, 128)	73856
MaxPooling2D	(None, 30, 30, 128)	0
Conv2D	(None, 28, 28, 128)	147584
MaxPooling2D	(None, 14, 14, 128)	0
Flatten	(None, 25088)	0
Dense	(None, 512)	12845568
Dense	(None, 1)	513
Total params: 13,086,913		
Test acc : 0.736		

Part B

모델 탐색 과정

Mobile phone Image data

VGG 사전 학습망과 이미지 보강

VGG사전 합성망 훈련 (인풋 150 × 150)		
Layer type	Output Shape	Param #
Conv2D	(None, 150, 150, 64)	1792
Conv2D	(None, 150, 150, 64)	36928
MaxPooling2D	(None, 75, 75, 64)	0
Conv2D	(None, 75, 75, 128)	73856
Conv2D	(None, 75, 75, 128)	147584
MaxPooling2D	(None, 37, 37, 128)	0
Conv2D	(None, 37, 37, 256)	295168
Conv2D	(None, 37, 37, 256)	590080
Conv2D	(None, 37, 37, 256)	590080
MaxPooling2D	(None, 18, 18, 256)	0
Conv2D	(None, 18, 18, 512)	1180160
Conv2D	(None, 18, 18, 512)	2359808
Conv2D	(None, 18, 18, 512)	2359808
MaxPooling2D	(None, 9, 9, 512)	0
Conv2D	(None, 9, 9, 512)	2359808
Conv2D	(None, 9, 9, 512)	2359808
Conv2D	(None, 9, 9, 512)	2359808
MaxPooling2D	(None, 4, 4, 512)	0
Flatten	(None, 8192)	14714688
Dense	(None, 256)	2097408
Dense	(None, 1)	257

Test acc : 0.851 / 이미지 보강시 0.842

VGG사전 합성망 훈련 (인풋 255 × 255)		
Layer type	Output Shape	Param #
Conv2D	(None, 255, 255, 64)	1792
Conv2D	(None, 255, 255, 64)	36928
MaxPooling2D	(None, 127, 127, 64)	0
Conv2D	(None, 127, 127, 128)	73856
Conv2D	(None, 127, 127, 128)	147584
MaxPooling2D	(None, 63, 63, 128)	0
Conv2D	(None, 63, 63, 256)	295168
Conv2D	(None, 63, 63, 256)	590080
Conv2D	(None, 63, 63, 256)	590080
MaxPooling2D	(None, 31, 31, 256)	0
Conv2D	(None, 31, 31, 512)	1180160
Conv2D	(None, 31, 31, 512)	2359808
Conv2D	(None, 31, 31, 512)	2359808
MaxPooling2D	(None, 15, 15, 512)	0
Conv2D	(None, 15, 15, 512)	2359808
Conv2D	(None, 15, 15, 512)	2359808
Conv2D	(None, 15, 15, 512)	2359808
MaxPooling2D	(None, 7, 7, 512)	0
Flatten	(None, 25088)	14714688
Dense	(None, 256)	6422784
Dense	(None, 1)	257

Test acc : 0.850

Drop out 추가시
test acc : 0.848

Part B

최종 모델

Mobile phone Image data

	간단 합성망 150	간단 합성망 255	간단 합성망 150 + 이미지 보강	VGG 사전학습망 150	VGG 사전학습망 255	VGG 255 + dropout	VGG 150 + 이미지 보강
Test acc	0.730	0.736	0.665	0.851	0.850	0.848	0.842
Epoch당 평균 시간	33초	166초	70초	265초	847초	757초	281초

감사합니다