

캐글을 활용한

회귀 / 분류 예측 분석



Table of Contents

목차



1 프로젝트 소개

3 모델 선택과 학습

5 향후 계획 및 개선 방향

2 데이터 이해와 전처리

4 성능 평가와 결과 분석

6 Q & A

01

프로젝트 소개

프로젝트 목적

kaggle

데이터
분석

+

리더보드
상위 10%

+

모델 이해

- 데이터 분석 및 머신러닝의 WorkFlow의 이해
- 캐글 리더보드 상위 10%를 목표로 계획과 전략 수립
- 모델에 대한 공부와 이해

분류 대회



생체신호를 이용한 흡연자 상태의 이진 분류 예측

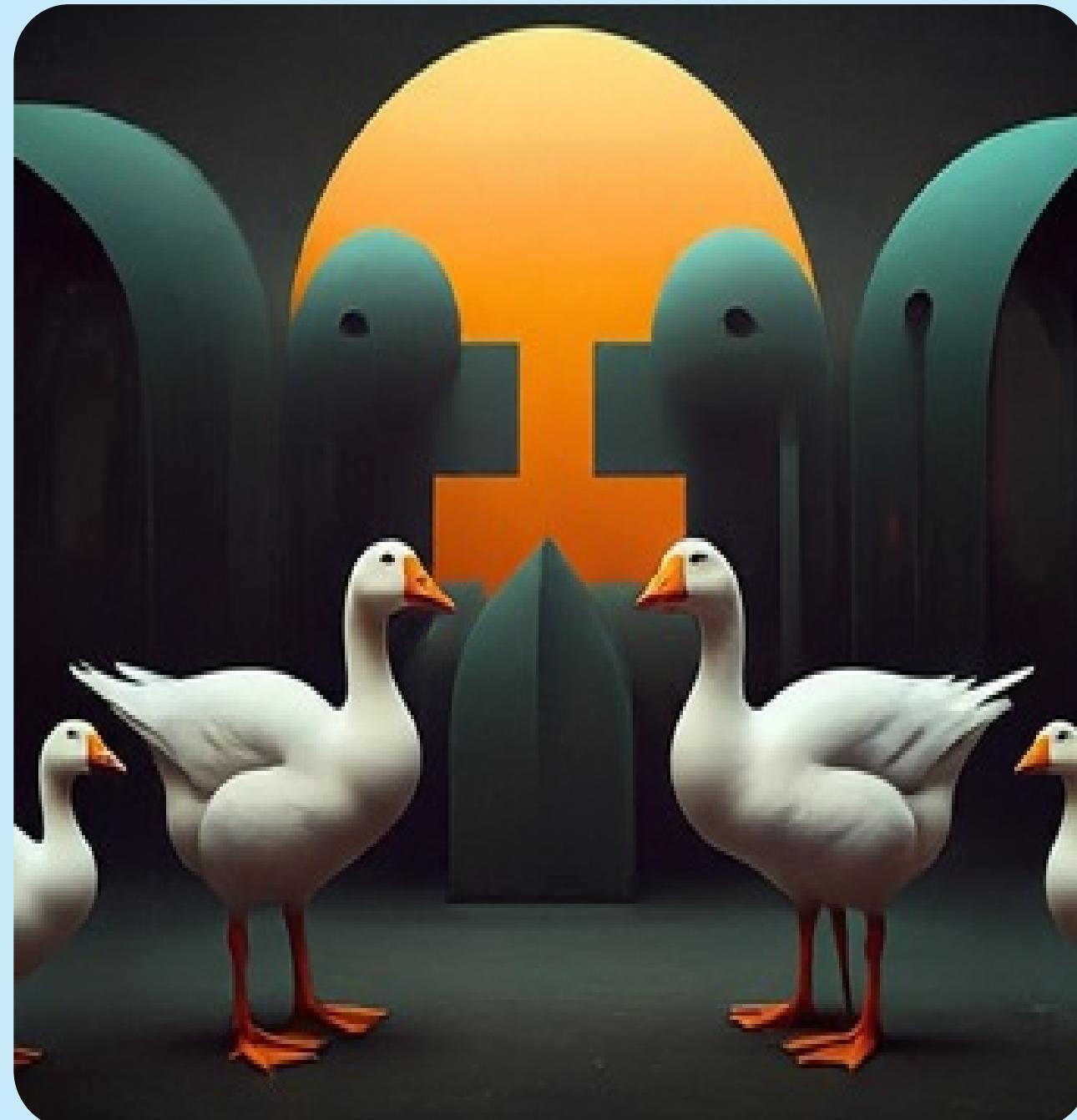
대회 목적

- 다양한 건강 지표에 대한 정보를 바탕으로 환자의 흡연 상태를 예측

평가지표

- ROC 곡선 아래 영역 (ROC_AUC_SCORE)

회귀 대회



전복 데이터를 활용한 나이 예측

대회 목적

- 다양한 물리적 측정을 통해 전복의 연령을 예측

평가지표

- RMSLE (Root Mean Squared Logarithmic Error)

분류

생체 신호를 이용한 흡연자 상태의 이진 예측

데이터 이해

기존
DATA

KAGGLE 제공
기본 데이터 SHAPE

- ✓ TRAIN : (159256, 24)
- ✓ TEST : (106171, 23)
- ✓ SUBMISSION : (106171, 23)

추가
DATA

추가적으로 수집한
데이터 SHAPE

- ✓ TRIAN_DATASET: (38984, 23)

데이터 이해

0	age	int64	0	18	85.0	20.0
1	height(cm)	int64	0	15	190.0	130.0
2	weight(kg)	int64	0	29	135.0	30.0
3	waist(cm)	float64	0	548	129.0	51.0
4	eyesight(left)	float64	0	20	9.9	0.1
5	eyesight(right)	float64	0	18	9.9	0.1
6	hearing(left)	int64	0	2	2.0	1.0
7	hearing(right)	int64	0	2	2.0	1.0
8	systolic	int64	0	128	233.0	71.0
9	relaxation	int64	0	94	146.0	40.0
10	fasting blood sugar	int64	0	259	423.0	46.0
11	Cholesterol	int64	0	279	445.0	55.0
12	triglyceride	int64	0	393	999.0	8.0
13	HDL	int64	0	123	359.0	4.0
14	LDL	int64	0	286	1860.0	1.0
15	hemoglobin	float64	0	144	21.1	4.9
16	Urine protein	int64	0	6	6.0	1.0
17	serum creatinine	float64	0	34	11.6	0.1
18	AST	int64	0	196	1090.0	6.0
19	ALT	int64	0	230	2914.0	1.0
20	Gtp	int64	0	444	999.0	2.0
21	dental caries	int64	0	2	1.0	0.0
22	smoking	int64	0	2	1.0	0.0

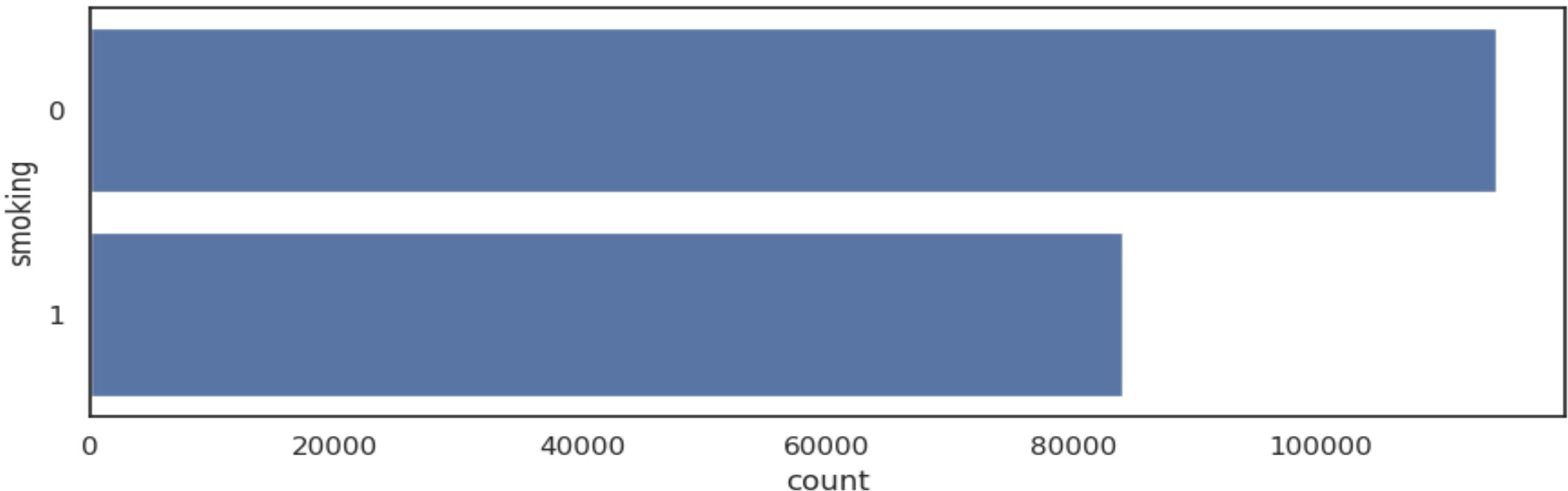
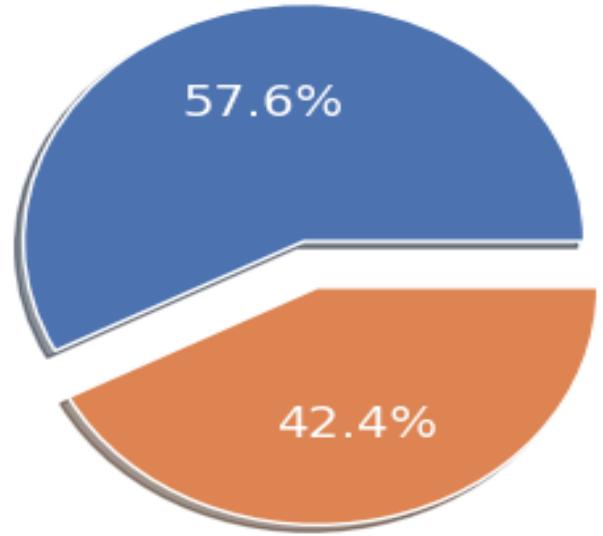
X피처: 22개
Y피처: 1개

- 나이 / 키 / 몸무게 등 기본적인 신체적인 피처
- 혈액 관련 피처
- 간 수치 관련 피처
- 시력 / 청력 피처
- Y피처 SMOKING은 0(흡연) / 1(비흡연)으로 이진 분류

EDA

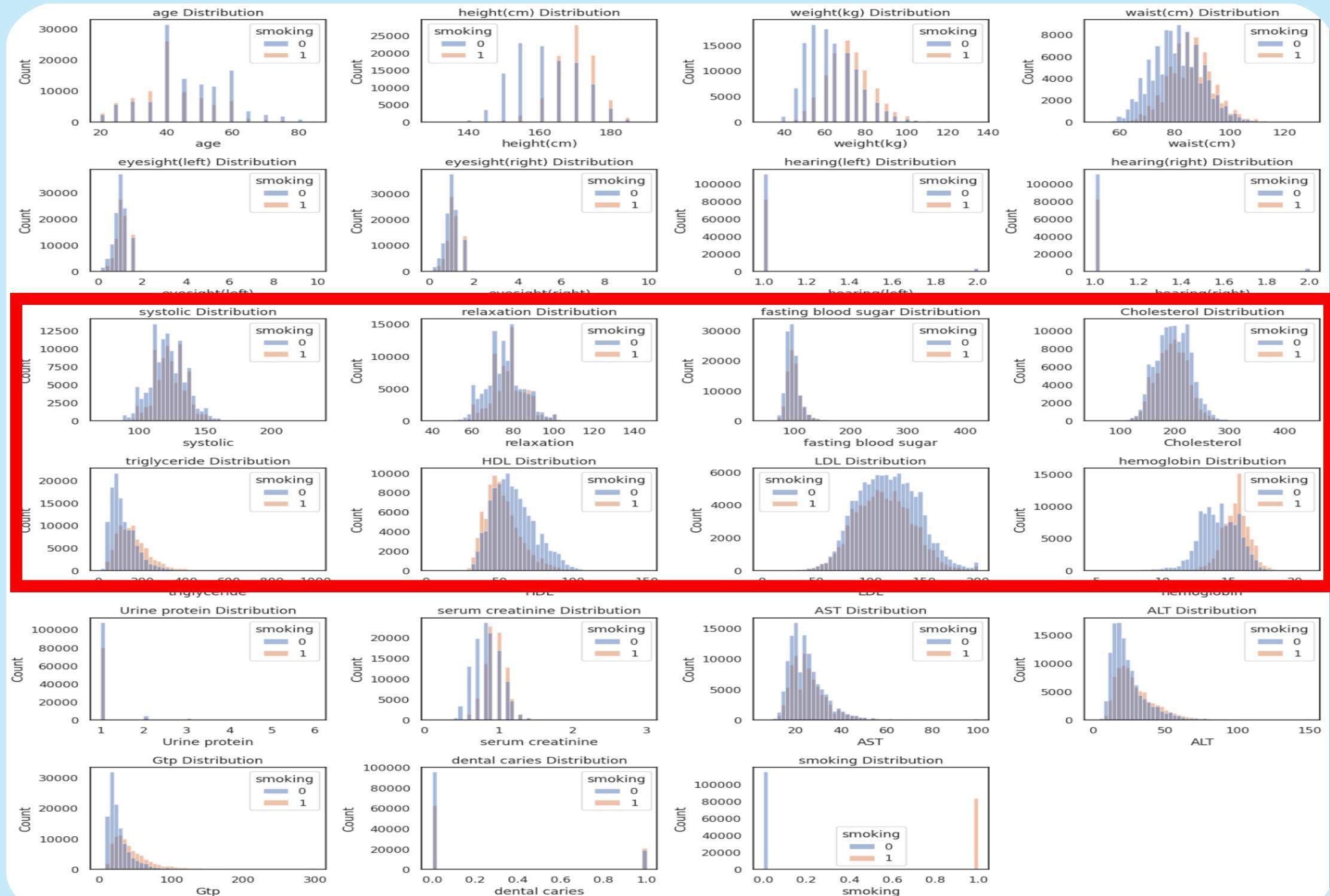
(타겟 분포 확인)

Smoking Distribution in Train



- 비흡연자 57.6 % / 흡연자 42.4%
- 균형을 이루는 타겟의 분포 확인

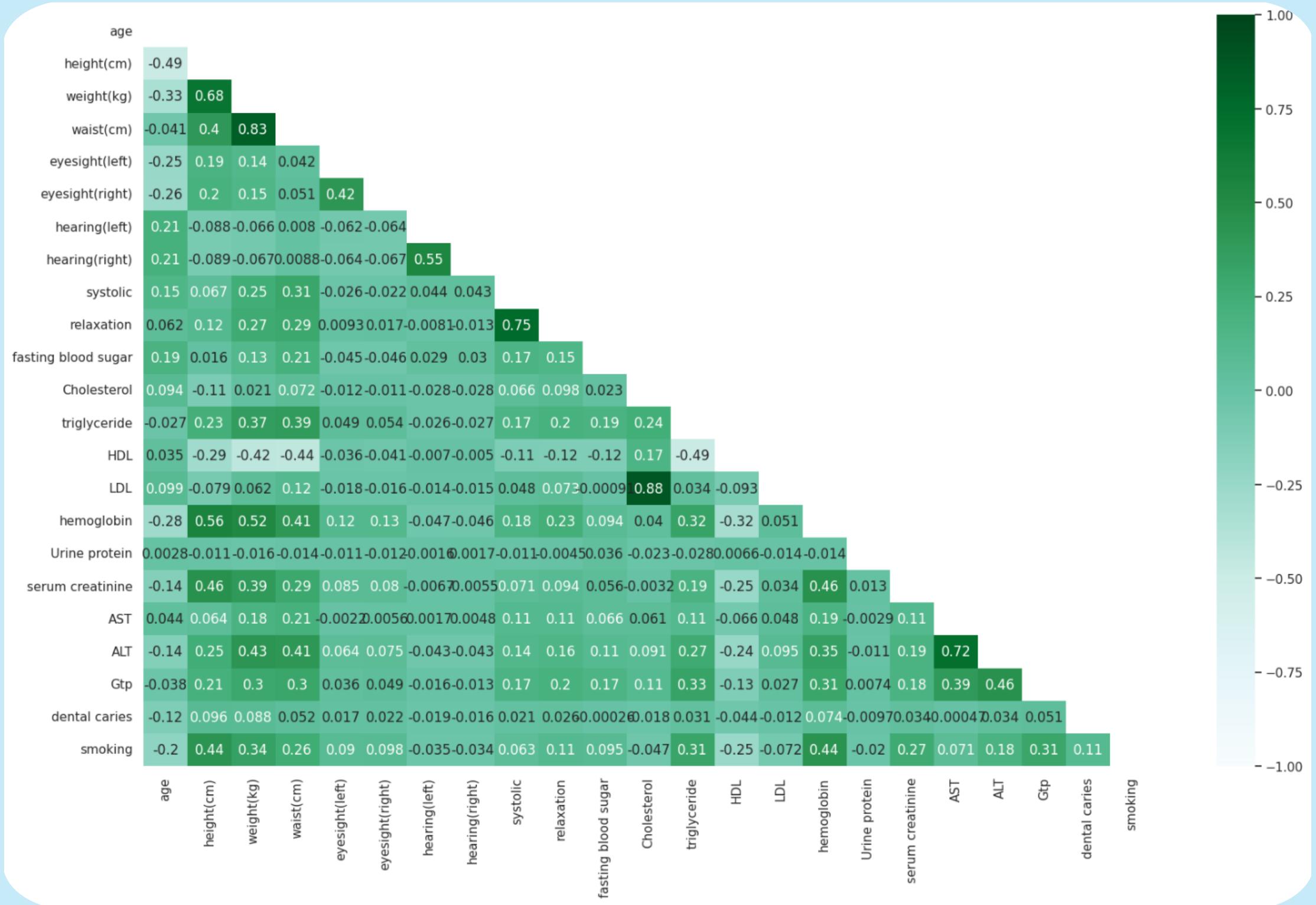
EDA (각 피처 분포 확인)



- 혈액 관련 피처들이 정규 분포의 형태를 따라감
- 다수 피처들이 오른쪽으로 꼬리가 긴 형태를 보임
- 이산형 피처 존재

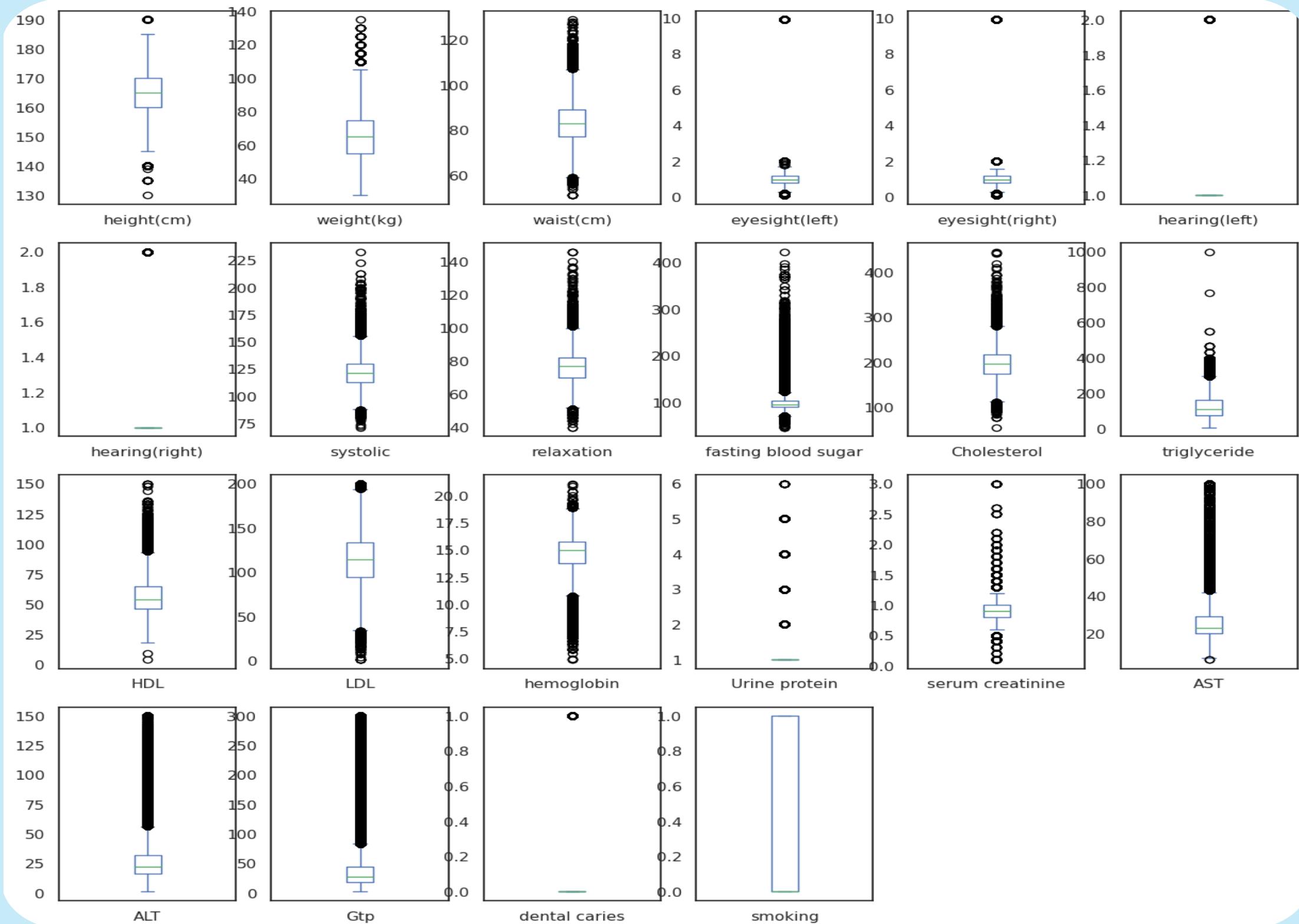
EDA

(상관관계 확인)



- 키 / 몸무게 / 허리둘레 피처에서 상관관계 확인
- 혈압과 콜레스테롤 피처에서 상관관계 확인
- 간 수치 관련한 피처에서 상관관계 확인

EDA (이상치 확인)

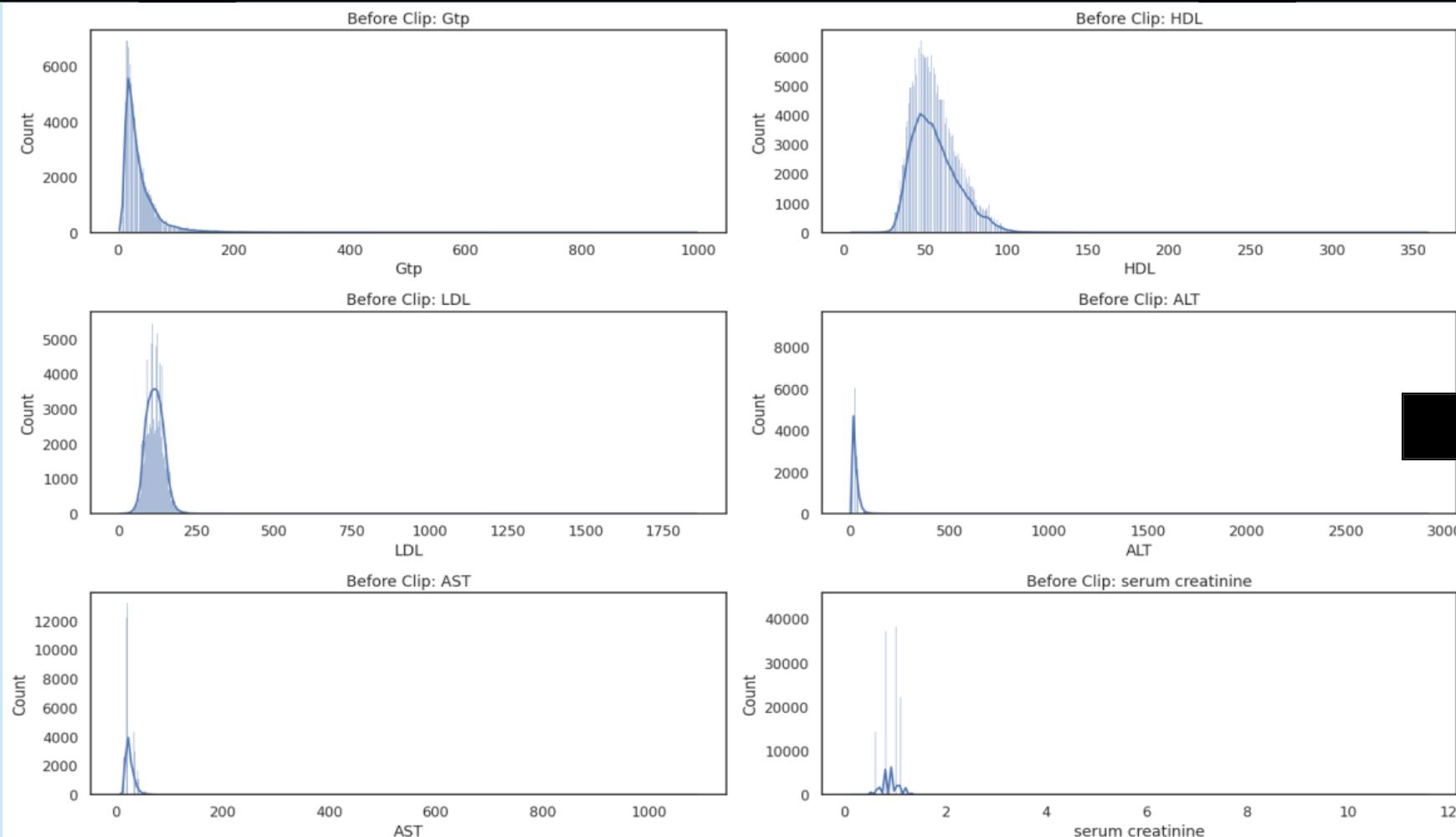


- 많은 피처에서 이상치 확인
- 이상치를 단순히 제거하거나
도메인적 지식을 바탕으로 제거

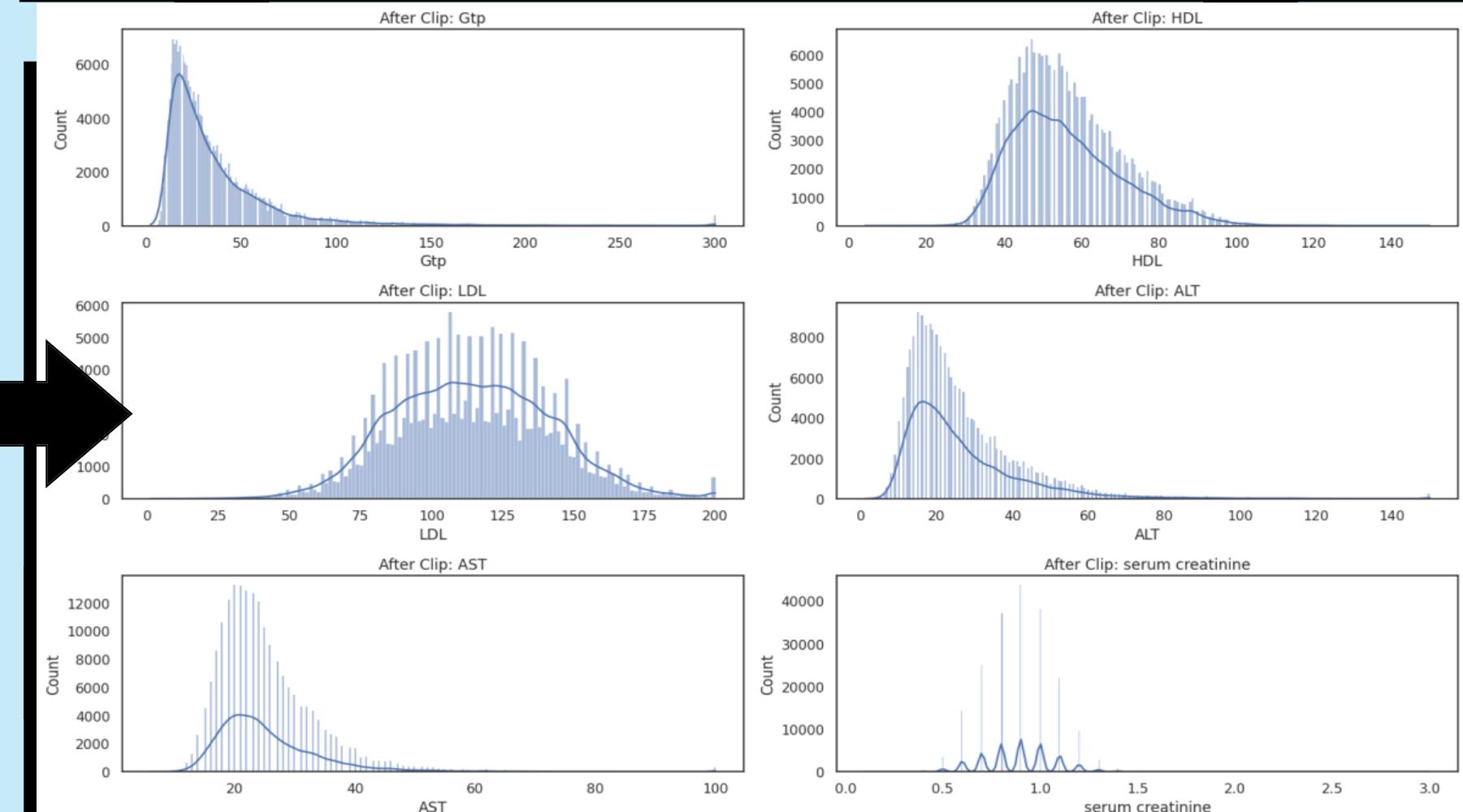
데이터 전처리

(특정 피처 이상치 조정)

이상치 조정 전



이상치 조정 후



- 오른쪽으로 긴 꼬리를 형성하는 분포
- 임계값을 설정하여 정규분포를 따르도록 조정

데이터 전처리

(특정 피처 이상치 조정)

```
1 # train 데이터에 대한  
2 train['Gtp'] = train['Gtp'].clip(lower = 0, upper = 300)  
3 train['HDL'] = train['HDL'].clip(lower = 0, upper = 150)  
4 train['LDL'] = train['LDL'].clip(lower = 0, upper = 200)  
5 train['ALT'] = train['ALT'].clip(lower = 0, upper = 150)  
6 train['AST'] = train['AST'].clip(lower = 0, upper = 100)  
7 train['serum creatinine'] = train['serum creatinine'].clip(lower = 0, upper = 3)  
8  
9 # test 데이터에 대한  
10 test['Gtp'] = test['Gtp'].clip(lower = 0, upper = 300)  
11 test['HDL'] = test['HDL'].clip(lower = 0, upper = 150)  
12 test['LDL'] = test['LDL'].clip(lower = 0, upper = 200)  
13 test['ALT'] = test['ALT'].clip(lower = 0, upper = 150)  
14 test['AST'] = test['AST'].clip(lower = 0, upper = 100)  
15 test['serum creatinine'] = test['serum creatinine'].clip(lower = 0, upper = 3)
```

- 특정 피처에 대하여 임계값을 통해 이상치 조정
- 의학적으로 인정하는 범위로 기준을 잡고 값을 피처의 분포를 조정

데이터 전처리

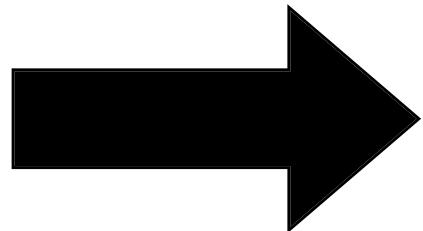
(병합 및 중복 제거)

```
1 train = pd.concat([train,train_dataset])
2 train
```

	age	height(cm)	weight(kg)	waist(cm)	eyesight(left)	eyesight(right)	hearing(left)	hearing(right)	systolic	relaxation	...	HDL	LDL	hemoglobin	Urine protein	serum creatinine	AST
0	55	165	60	81.0	0.5	0.6	1	1	135	87	...	40	75	16.5	1	1.0	22
1	70	165	65	89.0	0.6	0.7	2	2	146	83	...	57	126	16.2	1	1.1	27
2	20	170	75	81.0	0.4	0.5	1	1	118	75	...	45	93	17.4	1	0.8	27
3	35	180	95	105.0	1.5	1.2	1	1	131	88	...	38	102	15.9	1	1.0	20
4	30	165	60	80.5	1.5	1.0	1	1	121	76	...	44	93	15.4	1	0.8	19
...
38979	40	165	60	80.0	0.4	0.6	1	1	107	60	...	61	72	12.3	1	0.5	18
38980	45	155	55	75.0	1.5	1.2	1	1	126	72	...	76	131	12.5	2	0.6	23
38981	40	170	105	124.0	0.6	0.5	1	1	141	85	...	48	138	17.1	1	0.8	24
38982	40	160	55	75.0	1.5	1.5	1	1	95	69	...	79	116	12.0	1	0.6	24
38983	55	175	60	81.1	1.0	1.0	1	1	114	66	...	64	137	13.9	1	1.0	18

198240 rows × 23 columns

TRAIN.SHAPEx = (159256, 23) + (38984, 23)
TEST.SHAPEx = (106171, 22)



TRAIN.SHAPEx = (192723, 23)
TEST.SHAPEx = (106171, 22)

모델 선택과 학습

(계획 수립)

01

다양한 모델 및 기법

- LightGBM
- Catboost
- XGBoost
- 양상블

02

교차검증

- Fold 개수 조정

03

하이퍼 파라미터 조정

- GridSearchCV
- optuna
- 직접 파라미터 설정
- Kaggle discussion 참고

04

피처 엔지니어링

- 특정 피처에 임계값 설정
- 수치평 피처 스케일링
- 범주형 피처 인코딩
- 파생변수 생성

모델 선택과 학습

(모델 선택)

	(0.86803) submission_5Fold_depth(12)
	(0.86907) submission_stacker_robust
	(0.87041) submission_stacker_standard
	(0.87050) submission_stacker_minmax
	(0.87194)submission_5Fold(2)(임계값 -50)
	(0.87386) submission_5Fold(X)
	(0.87413)submission_5Fold(3)(임계값 +50)
	(0.87413)submission_5Fold
	(0.87474) submission_LGB
	(0.87553) submission_LGB_핫코딩_minmax(rate=0.3)
	(0.87626) submission_LGB(X)
	(0.87760) submission_LGB_핫코딩_minmax(rate=0.12)
	(0.87767) submission_LGB_핫코딩
	(0.87770) submission_LGB_핫코딩_minmax(boosting=gdbt)
	(0.87784) submission_LGB_fold_원핫인코딩
	(0.87795) lgb_submission_minmax
	(0.87797) lgb_submission_minmax
	(0.87802) lab_submission_minmax

성능 향상을 위한 여러가지 기법을 사용

- 양상분
- 스태킹

다양한 모델 사용

- LightGBM
- CatBoost
- XGBoost

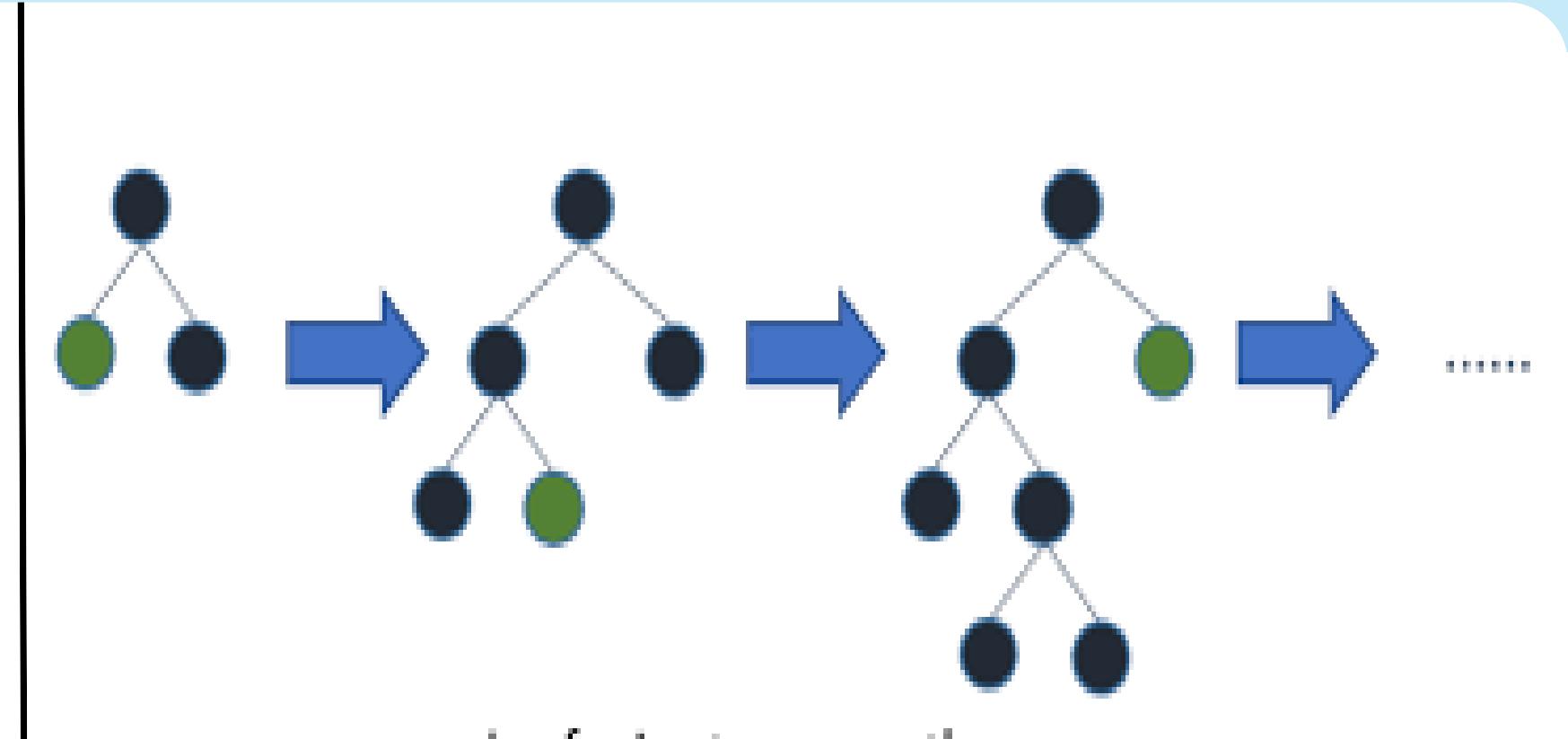
LightGBM 단일 모델이 가장 좋은 성능을 보이고 현재 데이터 적합하다고 판단

모델 선택과 학습

(선택된 모델)



LightGBM



- 그래디언트 부스팅
- 여러개의 약한 학습기 결합 후 강력한 학습기를 만드는 기법
- 대용량 데이터셋에 대한 빠른 속도와 높은 성능

- 리프 중심 트리 분할
- 효율적인 메모리 사용
- 다양한 손실 함수 지원

모델 선택과 학습

(모델 학습)

Input Data (22개)

범주형 피처	수치형 피처
hearing(left) hearing(right) Urine protein dental caries	age height(cm) weight(kg) waist(cm) eyesight(left) eyesight(right) systolic relaxation fasting blood sugar cholesterol triglyceride HDL / LDL hemoglobin serum creatinine AST / ALT / Gtp

Parameter

objective	"binary"
lambda_l1	3.1422772805786794
lambda_l2	0.3912821668022086
boosting_type	"gbdt"
num_leaves	120
feature_fraction	0.24955364718656628
bagging_fraction	0.9712078900905778
bagging_freq	9
min_child_samples	200
num_boost_round	600
learning_rate	0.1

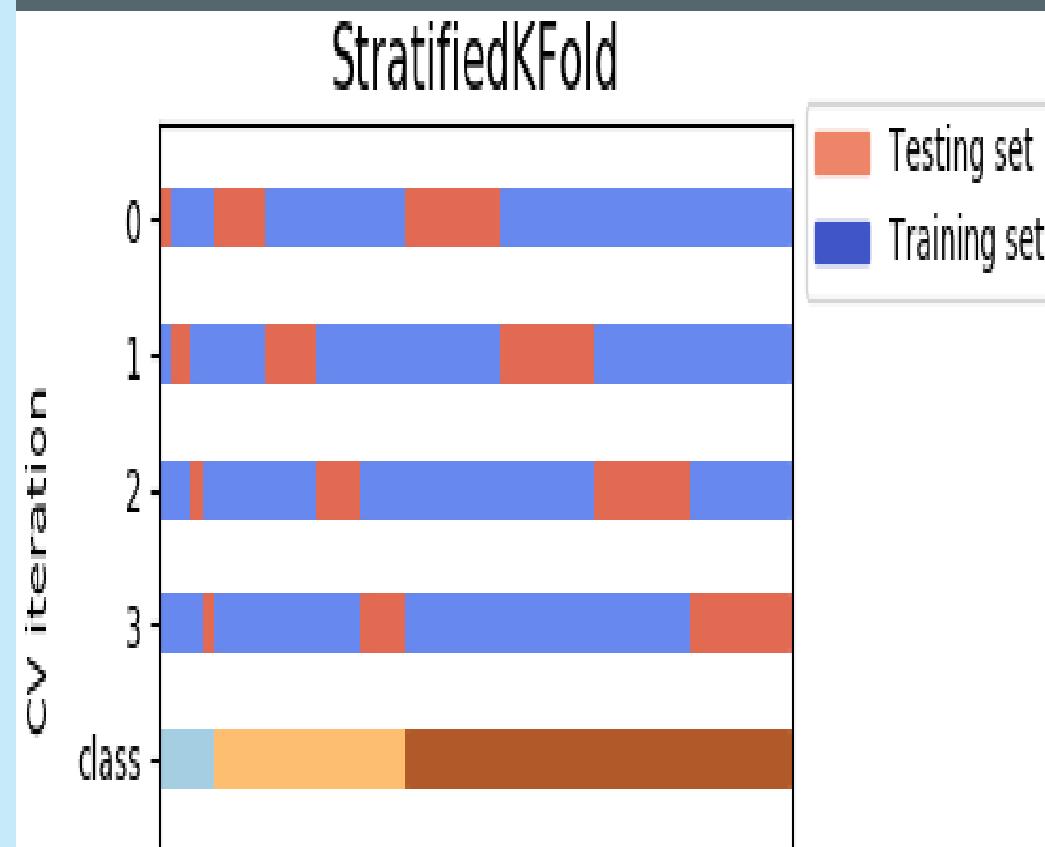


- 스케일링과 인코딩은 적용하지 않음
- GridSearchCV / optuna / kaggle discussion 활용하여 현재 모델에 최적의 파라미터를 발견

모델 선택과 학습

(성능 향상 및 안정성 평가)

교차검증 진행



- 모델의 안정성 및 성능 향상을 위해 교차 검증 진행
- 동일한 기준으로 split하기 위해 각 fold 비율 점검

Random_seed 변경

데이터:0, 난수 값: 1, auc: 0.8788207112082365
데이터:0, 난수 값: 3333, auc: 0.8787190682445074
데이터:0, 난수 값: 6666, auc: 0.8789924941491973
데이터:0, 난수 값: 9999, auc: 0.8787495958589154
데이터:1, 난수 값: 1, auc: 0.8721568414252926
데이터:1, 난수 값: 3333, auc: 0.8716565215032472
데이터:1, 난수 값: 6666, auc: 0.8718754542374207
데이터:1, 난수 값: 9999, auc: 0.8724189816110394
데이터:2, 난수 값: 1, auc: 0.8730651630670978
데이터:2, 난수 값: 3333, auc: 0.8741566657153299
데이터:2, 난수 값: 6666, auc: 0.8745638810094258
데이터:2, 난수 값: 9999, auc: 0.8736001010383352

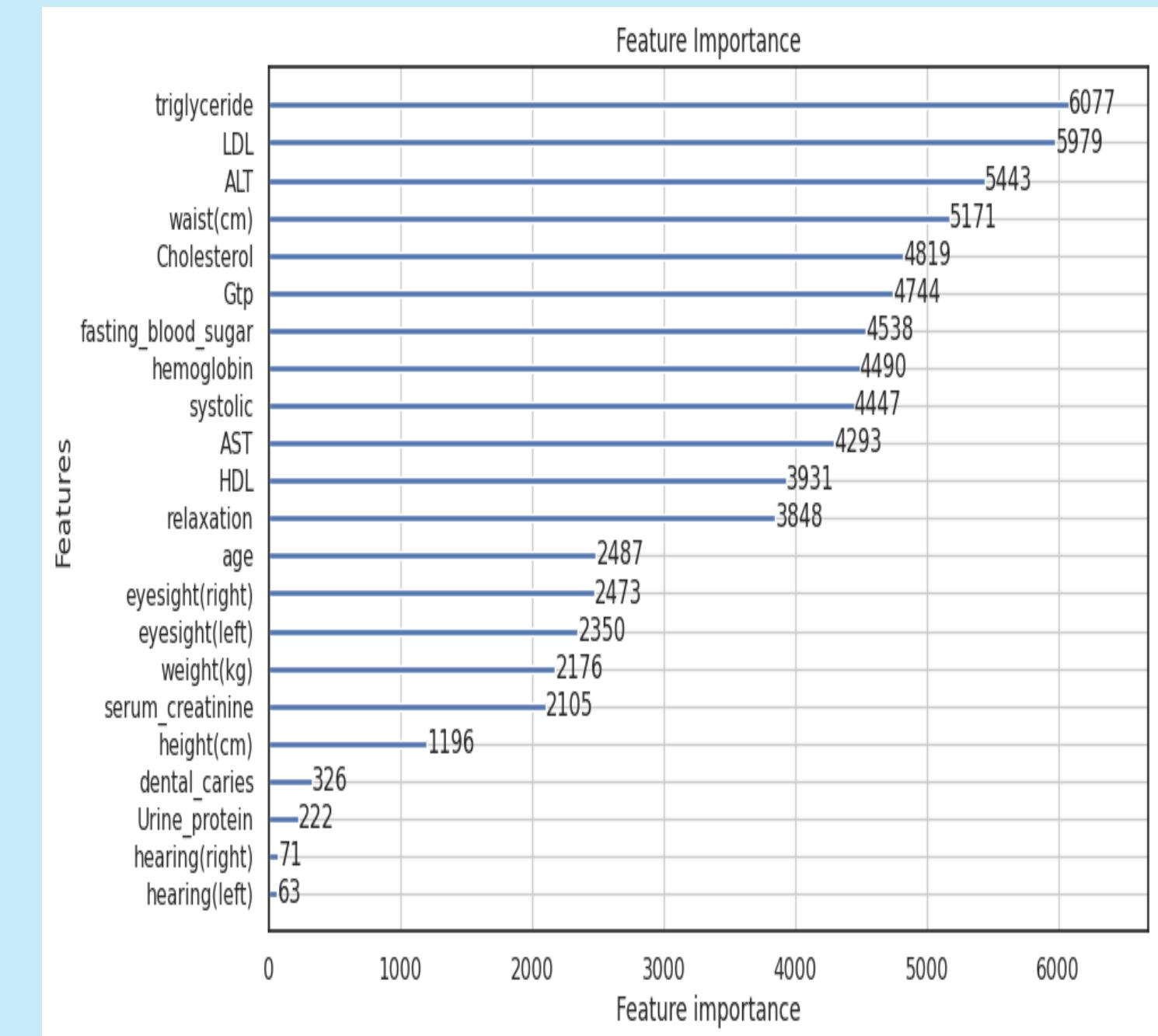
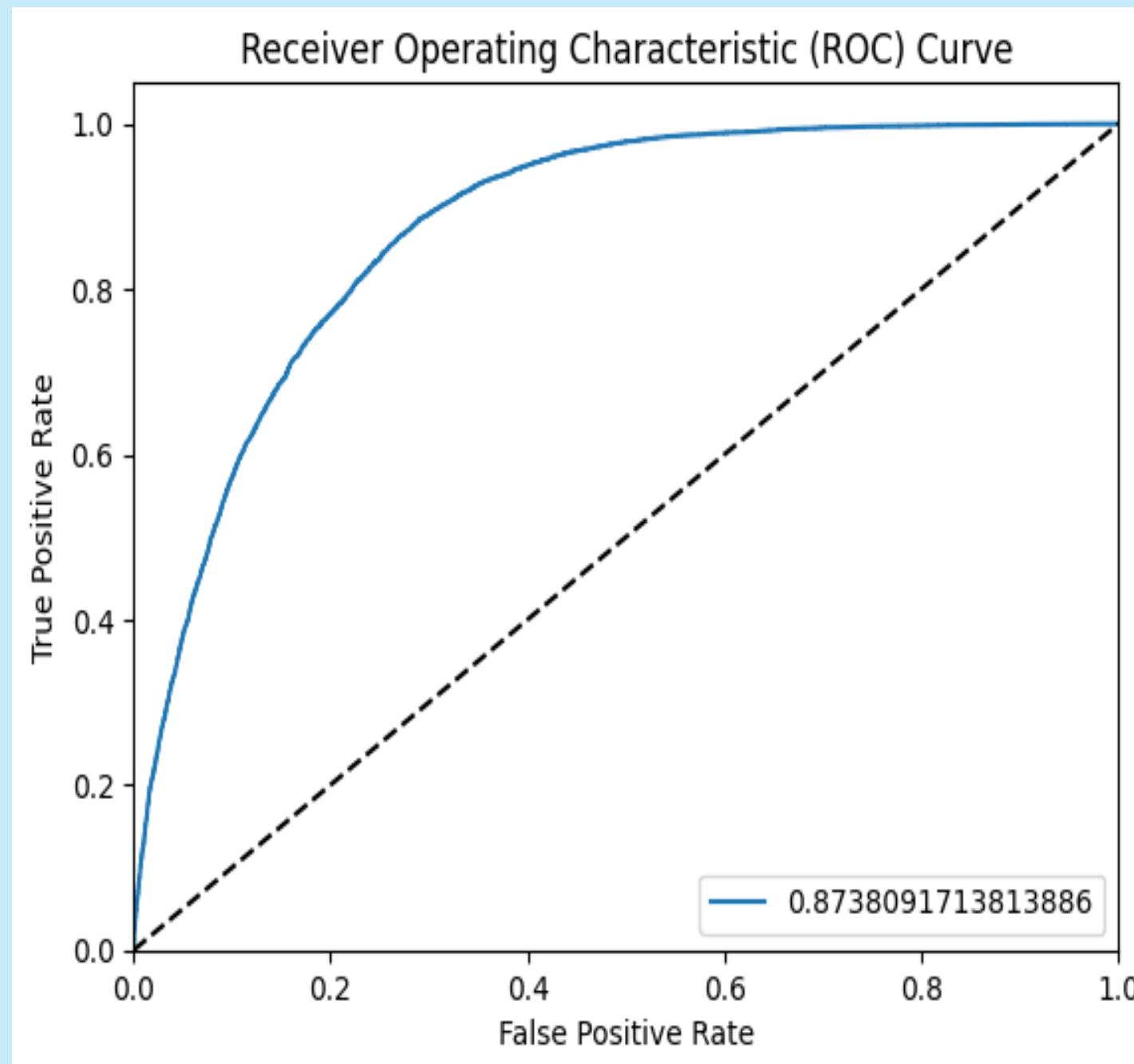
- 모델의 random_seed를 변경하여 안정성 평가

최종 제출 파일 생성

i	d	smok i ng
0	159256	0.660951
1	159257	0.210939
2	159258	0.335408
3	159259	0.016674
4	159260	0.659832

- OOB 방식을 활용하여 최종 제출 DATA 생성

성능평가 및 결과분석



- AUC 0.87로 좋은 성능을 가진 모델이라고 판단

- 혈액 / 혈관 및 콜레스테롤 수치가 상위권 차지
- 시력 피처를 제거해본 결과 성능이 감소하여 모델에 영향을 미치는 피처라는 것을 확인

성능평가 및 결과분석

(KAGGLE 제출 결과)

lgb_submission_final_8.csv				0.87805		
187	▼ 27	산		0.87806	5	5개월
188	▲ 5	완웨와뉴에		0.87806	5	6개월
189	▲ 5	에티엔A		0.87806	삼	5개월
190	▲ 5	슈밤 차반		0.87806	5	5개월
191	▼ 11	낸시		0.87799	11	6개월
192	▲ 19	파이살 알스레이드		0.87798	삼	5개월
193	▲ 7	은행 계좌		0.87797	10	6개월
194	▲ 7	유티아오		0.87794	24	5개월

목표로 했던 상위 10% 달성

회귀

전복 데이터를 활용한 나이 예측

데이터 이해

기존
DATA

KAGGLE 제공
기본 데이터 SHAPE

- ✓ TRAIN : (90615, 9)
- ✓ TEST : (60411, 8)
- ✓ SUBMISSION : (60411, 2)

추가
DATA

추가적으로 수집한
데이터 SHAPE

- ✓ TRIAN_DATASET: (4177, 9)

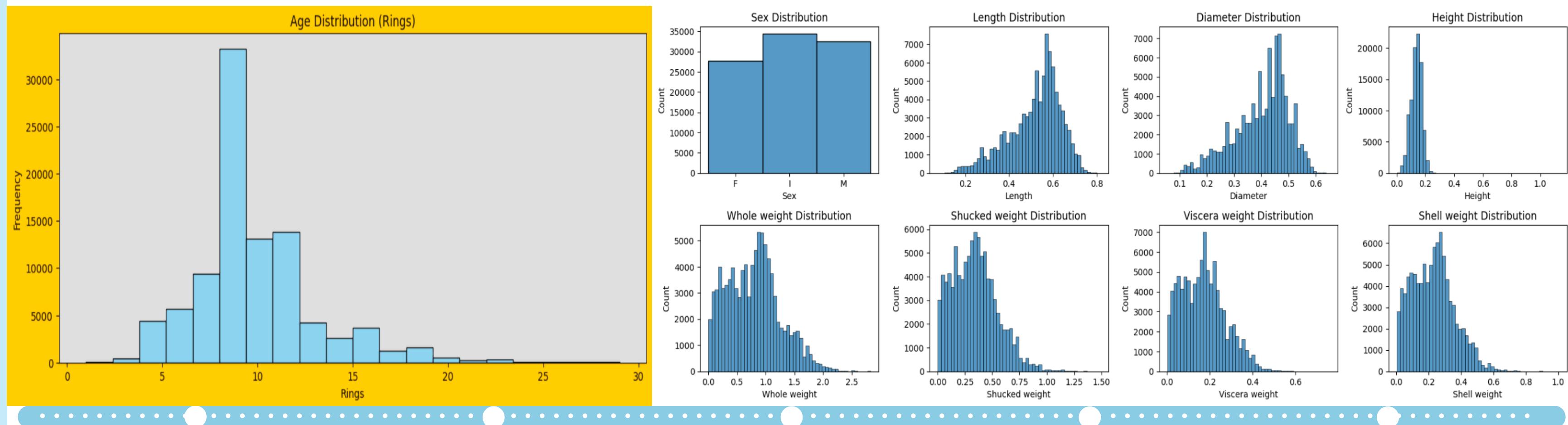
데이터 이해

feature	데이터 타입	결측값	고윳값	max	min
0 id	int64	0	90615	90614	0
1 Sex	object	0	3	M	F
2 Length	float64	0	157	0.815	0.075
3 Diameter	float64	0	126	0.65	0.055
4 Height	float64	0	90	1.13	0.0
5 Whole weight	float64	0	3175	2.8255	0.002
6 Whole weight.1	float64	0	1799	1.488	0.001
7 Whole weight.2	float64	0	979	0.76	0.0005
8 Shell weight	float64	0	1129	1.005	0.0015
9 Rings	int64	0	28	29	1

X피처: 8개
Y피처: 1개

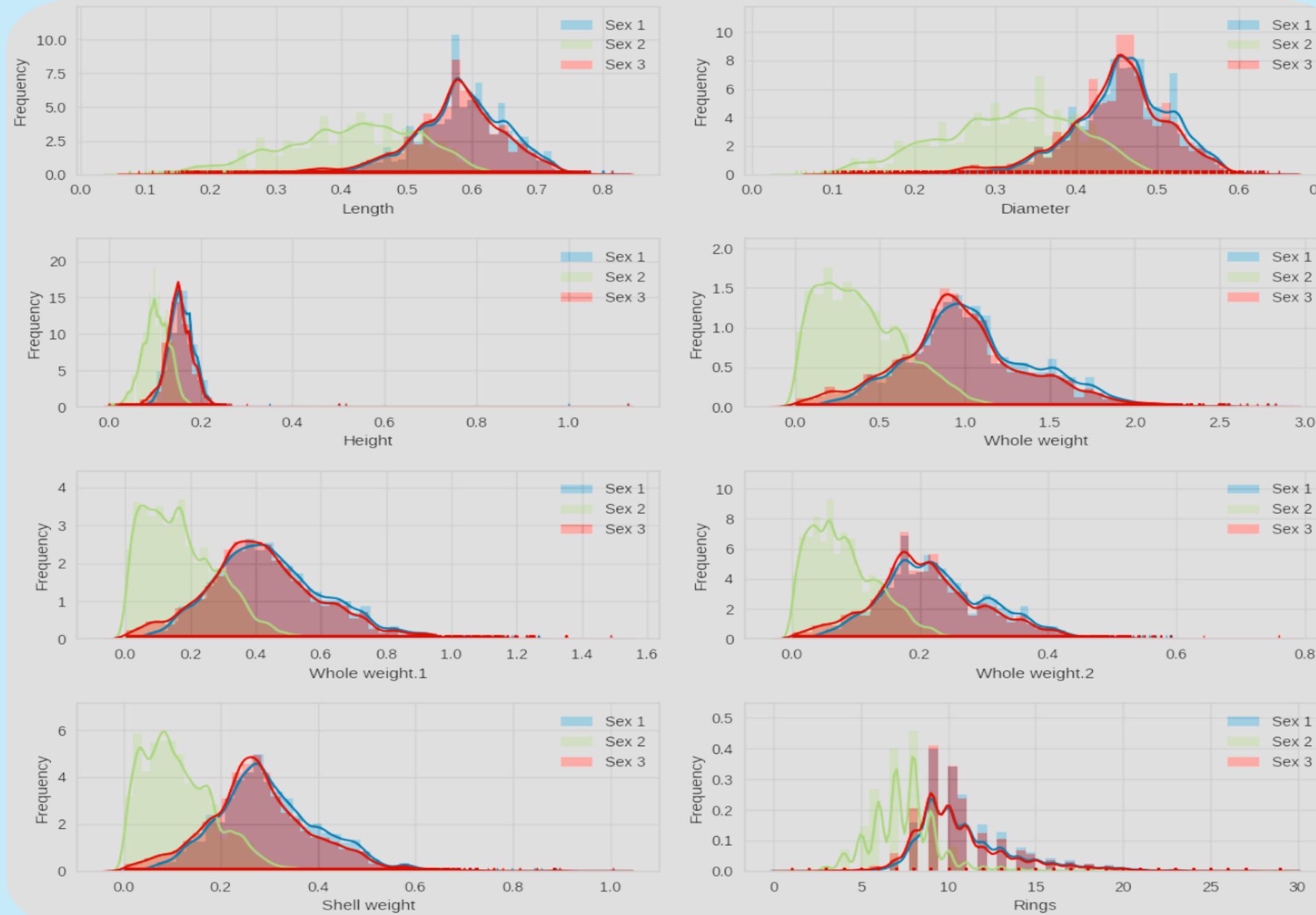
- 성별 / 무게 / 길이 / 높이
- 전복의 전체 무게 / 고기 무게 / 내장 무게 / 껍질 무게
- Y피처 Rings는 전복 나이를 나타내며 고윳값 1~29를 가짐

EDA (타겟 분포 확인)



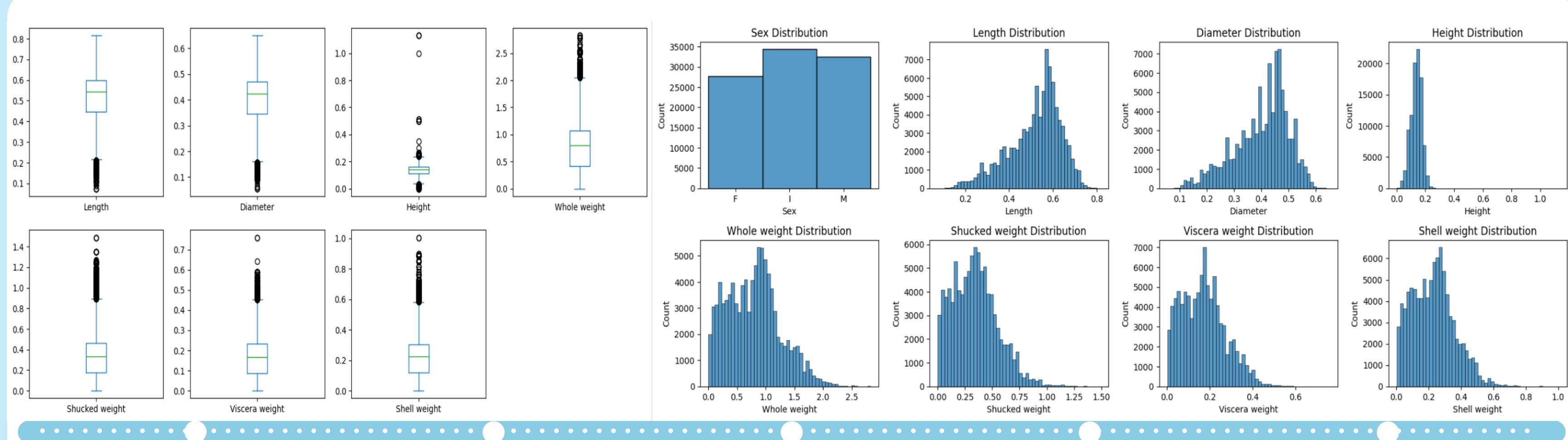
- 평균이 중앙값인 9.0을 상회하는 9.71 존재
- Y피처가 오른쪽으로 긴 꼬리를 형성
- 성별 / 길이 / 지름 피처를 제외하고 오른쪽으로 긴 꼬리를 형성

EDA (각 피처 분포 확인)



- **FEMALE과 MALE은 비슷한 분포**
- **IMMATURE는 다른 분포를 나타냄**

EDA (이상치 확인)

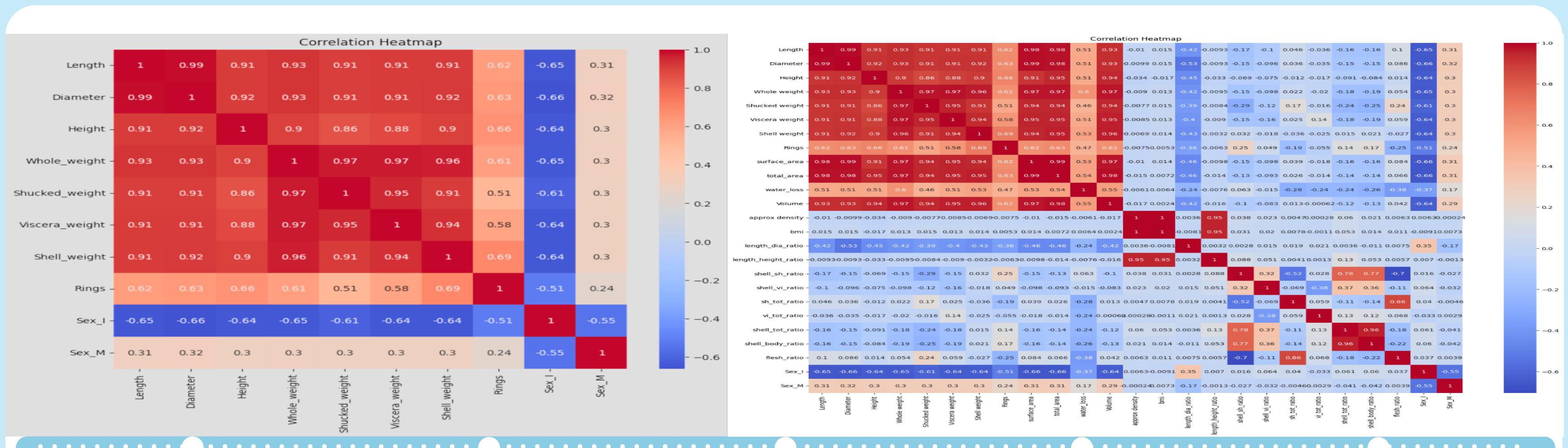


- 길이와 지름은 평균보다 낮은 값에서 이상치를 확인

- 나머지 피쳐들은 평균보다 높은 값에서 이상치를 확인

EDA

(상관관계 확인)



- 기존 데이터간의 강한 상관관계

데이터 전처리

(병합 및 중복 제거)

```

1 # train데이터와 orginal(abalone)데이터 병합
2 train = pd.concat([train, orginal], axis = 0, ignore_index=True)
3 train.head()

```

	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
0	F	0.550	0.430	0.150	0.7715	0.3285	0.1465	0.2400	11
1	F	0.630	0.490	0.145	1.1300	0.4580	0.2765	0.3200	11
2	I	0.160	0.110	0.025	0.0210	0.0055	0.0030	0.0050	6
3	M	0.595	0.475	0.150	0.9145	0.3755	0.2055	0.2500	10
4	I	0.555	0.425	0.130	0.7820	0.3695	0.1600	0.1975	9

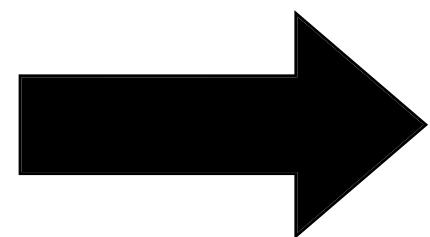
컬럼명 통일화

Feature
Sex
Length
Diameter
Height
Whole weight
Whole weight.1
Whole weight.2
Shell weight
Rings



Feature
Sex
Length
Diameter
Height
Whole weight
Shucked weight
Viscera weight
Shell weight
Rings

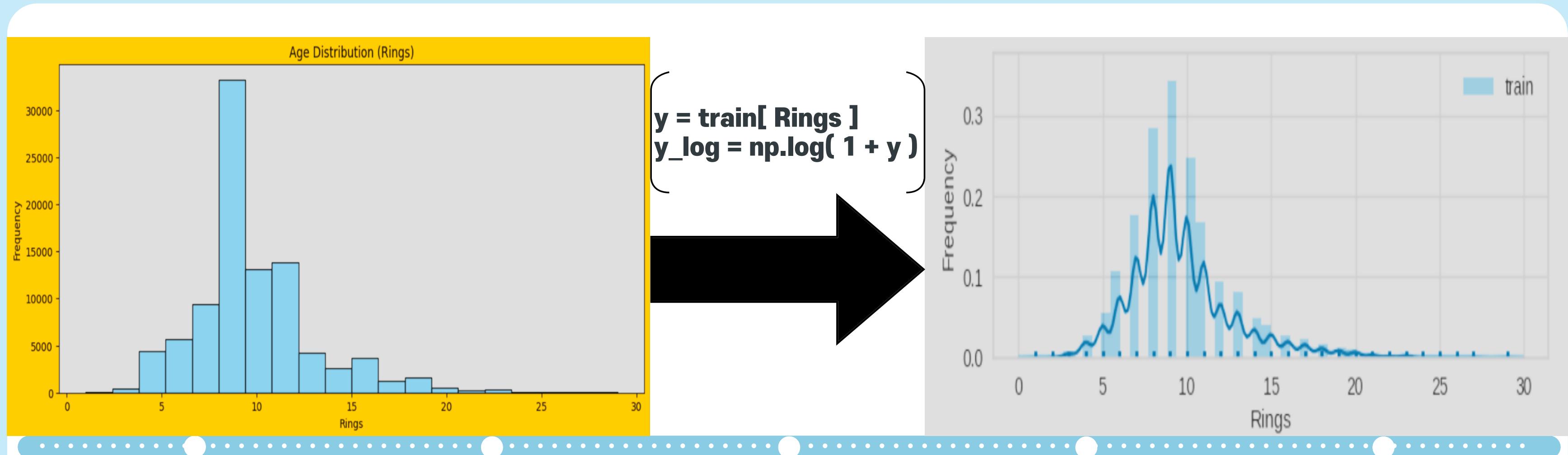
TRAIN.SHAPe = (90615, 9) + (4177, 9)
TEST.SHAPe = (60411, 8)



TRAIN.SHAPe = (94792, 9)
TEST.SHAPe = (60411, 8)

데이터 전처리

(타겟 로그 변환)

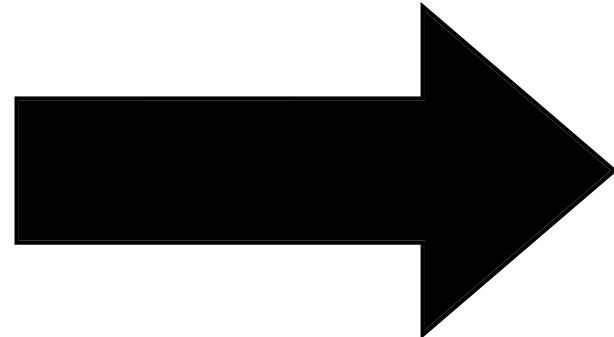


- 정규분포에 가깝게 타겟의 분포를 바꾸고, 오차에 대한 예측 오류 영향을 완화하기 위함
- 값이 0이거나 음수인 경우 -INF값이 발생하여 1을 더해 오류 방지

데이터 전처리

(성별 데이터 인코딩)

Sex
M
M
F
M
I
I



F	I	M
0	0	1
0	0	1
1	0	0
0	0	1
0	1	0
0	1	0

- 모델이 학습하기 쉽도록 범주형 피처에 ONEHOT 인코딩 처리
- 남 / 여 / 미성숙체를 구분하여 0, 1 형태로 처리

모델 선택과 학습

(계획 수립)

01

다양한 모델 및 기법

- LightGBM
- Catboost
- XGBoost
- 양상블

02

교차검증

- Fold 개수 조정

03

하이퍼 파라미터 조정

- GridSearchCV
- optuna
- 직접 파라미터 설정
- Kaggle discussion 참고

04

피처 엔지니어링

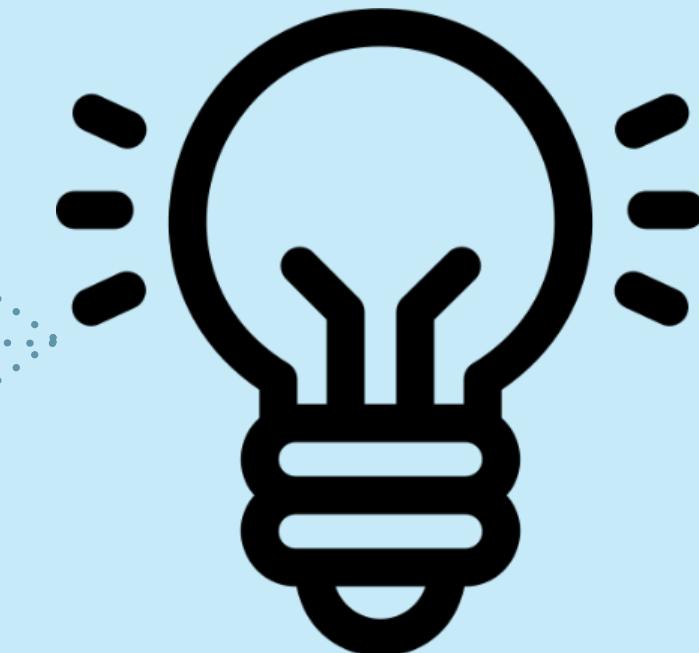
- 특정 피처에 임계값 설정
- 수치평 피처 스케일링
- 범주형 피처 인코딩
- 파생변수 생성

모델 선택과 학습

(선택된 모델)

Input Data (8개)

범주형 피처	수치형 피처
Sex	Length Diameter Height Whole weight Shucked weight Viscera weight Shell weight



- 기존 피처와 모델의 DEFAULT 파라미터를 통해 모델별 비교
- 파생변수 추가 후 성능확인 결과 성능 저하

모델 선택과 학습

(선택된 모델)

LIGHTGBM

num_leaves	200
bagging_freq	7
boosting_type	“gbdt”
min_child_samples	91
objective	‘regression’
learning_rate	0.0955
bagging_fraction	0.6502062728410578
feature_fraction	0.7058843944694884

XGBOOST

max_depth	20
booster	‘dart’
lambda	0.456836886068415
alpha	0.6422509164613671
subsample	0.9365423486036913
objective	‘reg:squaredlogerror’
learning_rate	0.0955
colsample_bytree	0.8111849113860014

CATBOOST

depth	20
max_bin	464
min_data_in_leaf	78
grow_policy	Lossguide
subsample	0.83862137638162
l2_leaf_reg	8.365422739510098
random_strength	3.296124856352495
learning_rate	0.0955

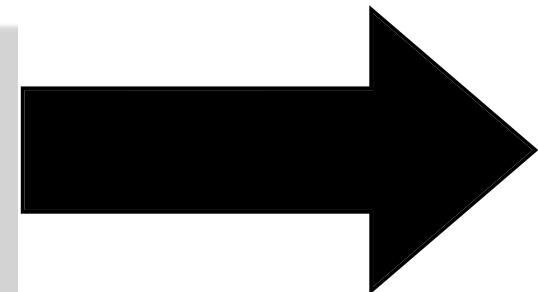
- OPTUNA / GRIDSEARCHCV / KAGGLE DISCUSSION 활용하여
최적의 파라미터 값 발견

모델 선택과 학습

(선택된 모델)

```
[1] top_3_model_gender = compare_models(fold=5, round=3, n_select=3, errors='ignore')
```

Model		MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	0.128	0.030	0.173	0.537	0.053	0.054	9.280
lightgbm	Light Gradient Boosting Machine	0.128	0.030	0.174	0.535	0.053	0.054	0.518
rf	Random Forest Regressor	0.131	0.031	0.177	0.520	0.054	0.055	50.704
xgboost	Extreme Gradient Boosting	0.130	0.031	0.177	0.520	0.054	0.055	1.172



```
[2] reg_blended_FM = blend_models(estimator_list=top_3_model_gender, fold=10)
```

	MAE	MSE	RMSE	R2	RMSLE	MAPE
Fold						
0	0.1279	0.0301	0.1734	0.5338	0.0527	0.0534
1	0.1290	0.0306	0.1750	0.5261	0.0531	0.0539
2	0.1297	0.0342	0.1850	0.5029	0.0614	0.0538
3	0.1300	0.0299	0.1730	0.5409	0.0512	0.0551
4	0.1249	0.0278	0.1667	0.5530	0.0485	0.0523
5	0.1263	0.0297	0.1722	0.5418	0.0529	0.0534
6	0.1267	0.0292	0.1708	0.5569	0.0510	0.0541
7	0.1270	0.0295	0.1717	0.5421	0.0529	0.0538
8	0.1245	0.0279	0.1669	0.5706	0.0518	0.0528
9	0.1285	0.0290	0.1702	0.5534	0.0495	0.0538
Mean	0.1274	0.0298	0.1725	0.5421	0.0525	0.0536
Std	0.0018	0.0017	0.0049	0.0177	0.0033	0.0007

- AUTOML을 통해 상위 TOP3모델 블렌딩 + 10FOLD 교차 검증 진행
- RMSLE 0.14700으로 상위 25%이므로 전략을 변경하여 VOTING으로 진행

성능평가 및 결과분석

(성능향상)

VOTING

- (0.14564) submission_voting_fold10(파라미터 수정1)
- (0.14565) submission_voting_fold10(파라미터 수정3)
- (0.14568) submission_voting_fold10(파라미터 수정2)
- (0.14576) submission_voting_fold5(원본 파라미터)
- (0.14606) submission_voting_feature(remains, volume)

- 파생 변수 사용시 성능 저하

VOTING + FOLD CHANGE

- (0.14559) submission_voting_fold11(최종)
- (0.14559) submission_voting_fold13(최종)
- (0.14559) submission_voting_fold14(최종)
- (0.14560) submission_voting_fold12(최종)
- (0.14563) submission_voting_fold15(최종)
- (0.14564) submission_voting_fold10(최종)
- (0.14565) submission_voting_fold9(최종)
- (0.14567) submission_voting_fold7(최종)
- (0.14567) submission_voting_fold8(최종)
- (0.14569) submission_voting_fold6(최종)
- (0.14576) submission_voting_fold5(최종)

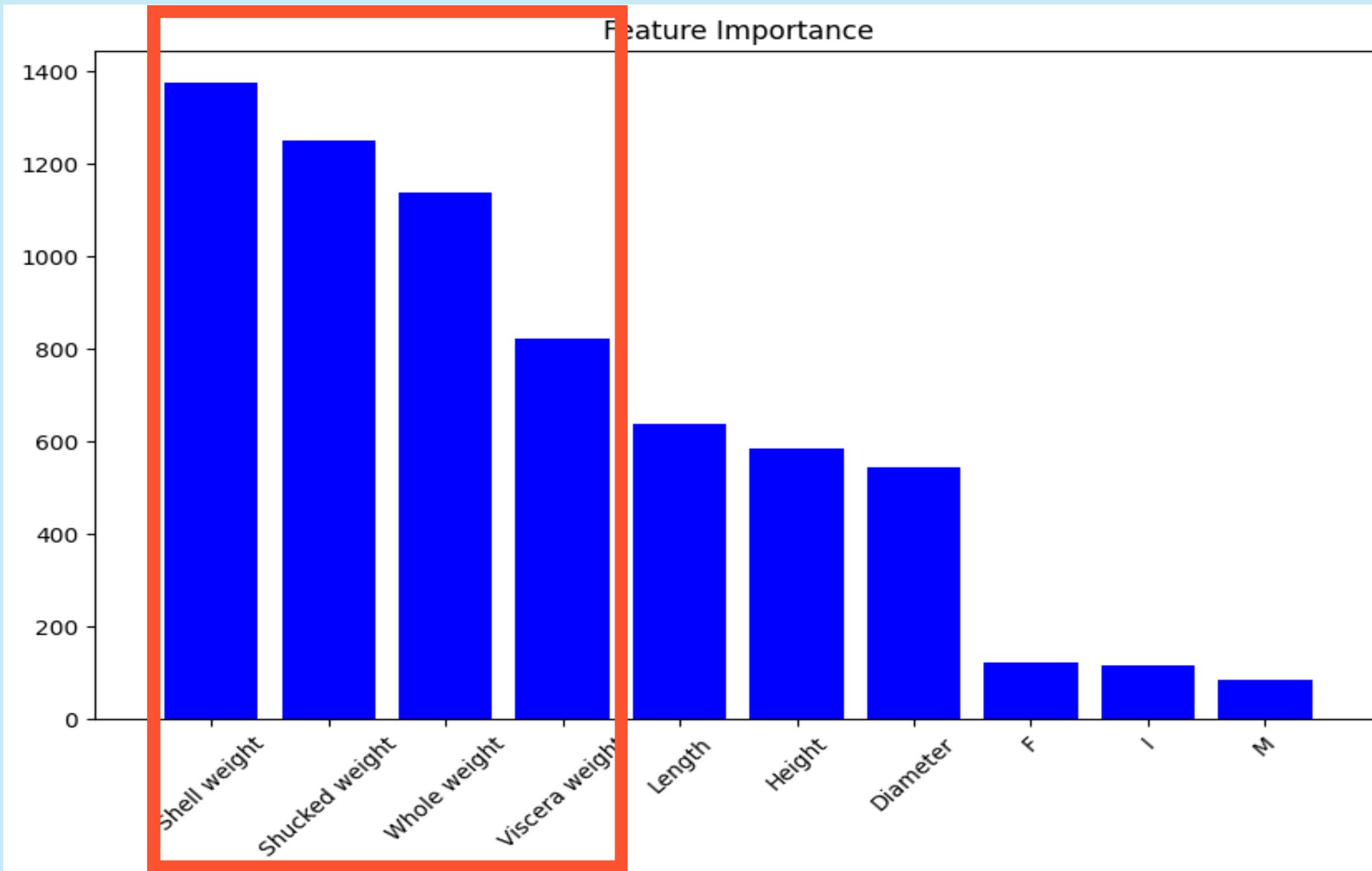
- FOLD 범위 5 ~ 15 확인
- 11 / 13 / 14 FOLD가 최적

VOTING + ENSEMBLE

- (0.14558) voting_fold10_top3(가중치_1, 1, 1)
- (0.14559) voting_fold10_top5(가중치_1, 1, 1, 1, 1)
- (0.14565) voting_fold10_top4(가중치_10, 10, 1, 1)_ensemble
- (0.14565) voting_fold10_top5(가중치 모두 1로 동일)_ensemble
- (0.14566) voting_fold10_top4(가중치_10, 1, 1, 1)_ensemble
- (0.14566) voting_fold10_top4(가중치_10, 10, 10, 1)_ensemble
- (0.14566) voting_fold10_top7(가중치_10, 9, 8, 7, 6, 5, 4)_ensemble

- 최적의 모델에 가중치를 적용하여 양상을 진행

성능평가 및 결과분석



- 무게 관련 피처가 중요도에서 높은 순위를 가짐

성능평가 및 결과분석

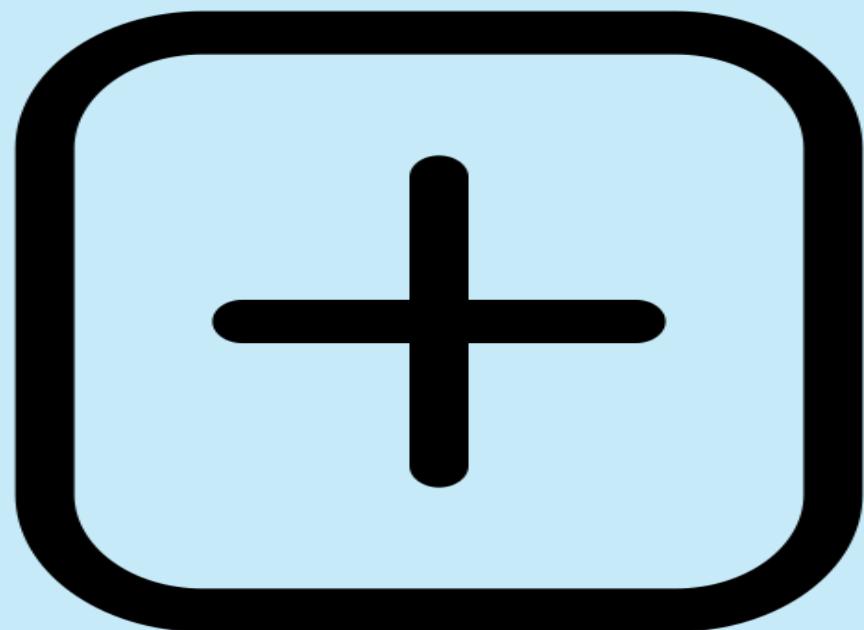
(KAGGLE 제출 결과)

vote_fold_top3(_111).csv
완료 · 5시간 전

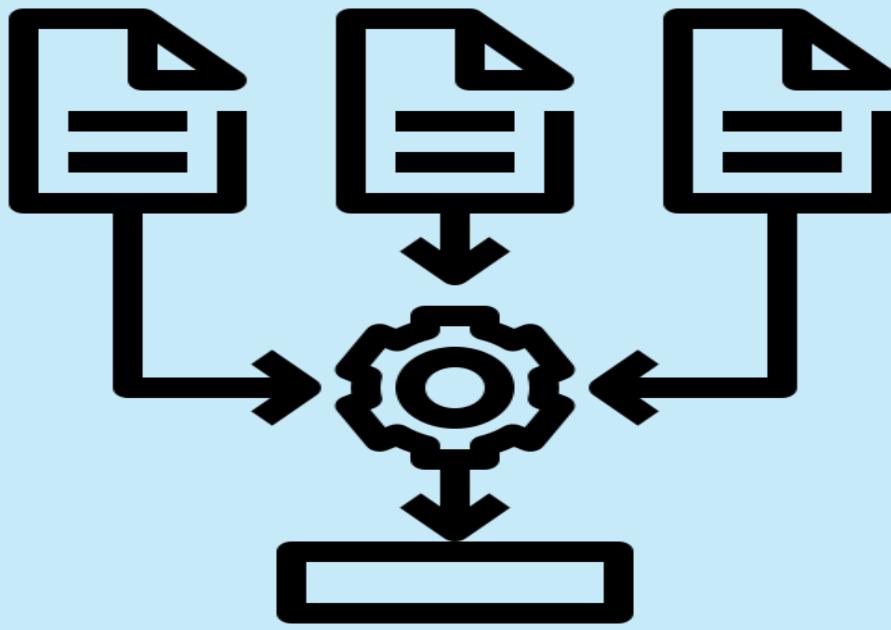
번호	제출자	평균 평가점수	제출일	상위 10%
199	하디 가루디	0.14557	삼 7일	
200	전사	0.14557	11 5d	
201	파벨 니콜라이체프	0.14557	1 12일	
202	아흐메드 아불크헤어	0.14557	1 10일	
203	크리스베밥	0.14557	53 1일	
204	곽재우	0.14558	16 1일	0.14558
 최고의 출품작입니다! 가장 최근 제출하신 점수는 0.14558로, 이전 점수인 0.14559보다 향상된 수치입니다. 잘 했어!				이것을 트윗하세요
205	시밤 싱	0.14558	1 14일	
206	데이터투	0.14558	5 8일	
207	조자성	0.14559	10 1일	
208	미즈틱	0.14559	17 1일	
209	GMROH637	0.14561	6 11일	

목표로 했던 상위 10% 달성

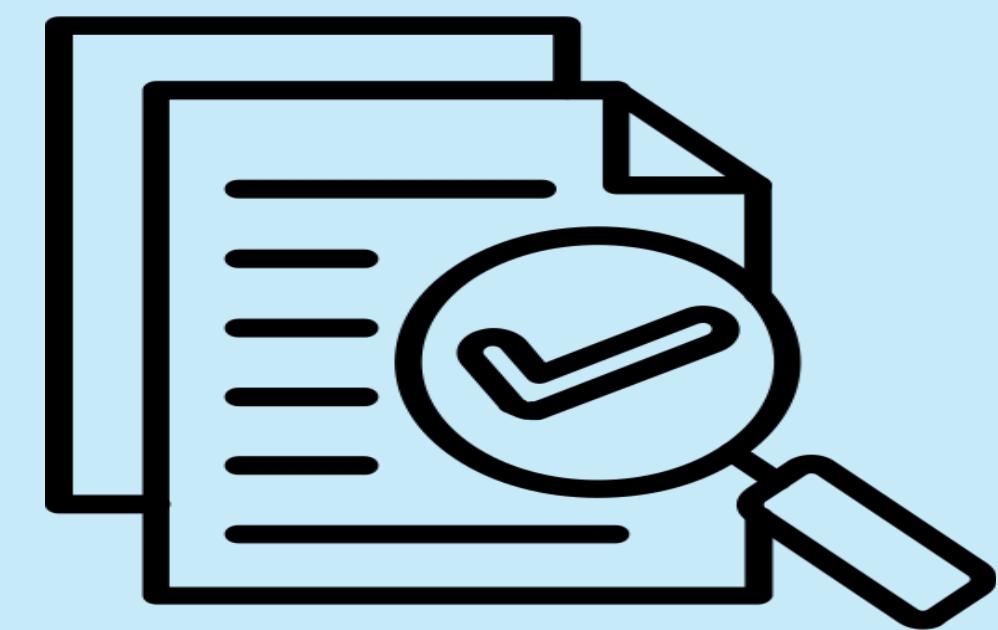
향후 계획 및 개선방향



- 추가적인 데이터 확보
- 새로운 파생변수 계획



- 다양한 양상을 모델 적용
- 딥러닝 모델 사용 고려



- 새로운 인사이트 발견

Q & A

출처

8-1. 참고자료

<https://github.com/Koda98/smoker-status-prediction/tree/main>
<https://www.kaggle.com/code/xxxxyyyy80008/smoker-status-prediction-lightgbm-baseline-no-fe>
<https://www.kaggle.com/code/arunklenin/ps3e24-smoking-cessation-prediction-binary>
<https://www.kaggle.com/code/arunklenin/ps4e4-abalone-age-prediction-regression/notebook#4.1-New-Features>
<https://www.kaggle.com/code/bunny11/voting-classifier/notebook>
<https://www.kaggle.com/code/satyaprakashshukl/cb-regression-analysis/notebook>

8-2. 출처

<https://www.kaggle.com/competitions/playground-series-s3e24/overview>
<https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction-using-biosignals>
<https://archive.ics.uci.edu/dataset/1/abalone>
<https://www.kaggle.com/competitions/playground-series-s4e4>

감사합니다