

Kaggle 다중 분류 모델 대회

팀장 : 박지건

팀원 : 김수현

팀원 : 양인선

팀원 : 송인동

Contents

1. Introduction to the competition
2. Evaluation metrics
3. Modeling Process


INTRODUCTION

E.D.A

**MODEL
EVALUATION**

1. RandomForest
2. XGBoost
3. CatBoost
4. LightGBM
5. Final Model Selection

1. Introduction to the competition


 KAGGLE · PLAYGROUND PREDICTION COMPETITION · 5 DAYS AGO

Late Submission

...

Multi-Class Prediction of Obesity Risk

Playground Series - Season 4, Episode 2



주제	비만 위험도 예측
유형	Playground
제출방식	Simple Competition (csv 파일 제출)
주최	Kaggle
문제 유형	멀티-클래스 분류(Multiclass classification)
데이터 타입	정형(Tabular) 데이터
평가지표	정확도(Accuracy)
대회 참가팀	3587팀
대회 기간	24.02.01 ~ 24.02.29, 11:59 PM UTC

대회 목표

개인의 다양한 생활 습관과 건강 데이터 등 다양한 요인들을 기반으로 개인의 비만 위험 수준을 예측하는 것

참가인원

4 Persosns

참여기간

7days (2.23~2.29)

1. Introduction to the competition

데이터셋 소개

- 대상: 멕시코, 페루, 콜롬비아 국가의 14세에서 61세 사이의 사람들
- 내용: 다양한 식습관과 신체 상태를 반영한 비만 수준 추정치
- 생성: 원본 비만 위험 데이터셋을 기반으로 심층 학습 모델 훈련을 통해 생성 -> 원본 데이터셋을 사용하면 성능이 향상

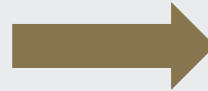
훈련 데이터 셋

데이터 수: 20758개
컬럼 수: 18개



원본 데이터 셋

데이터 수: 2111개
컬럼 수: 17개



결합된 데이터 셋


데이터 수: 22869개
컬럼 수: 17개 (id 컬럼 제외)

결측치와 중복되는 값은 없음

2. Evaluation metrics

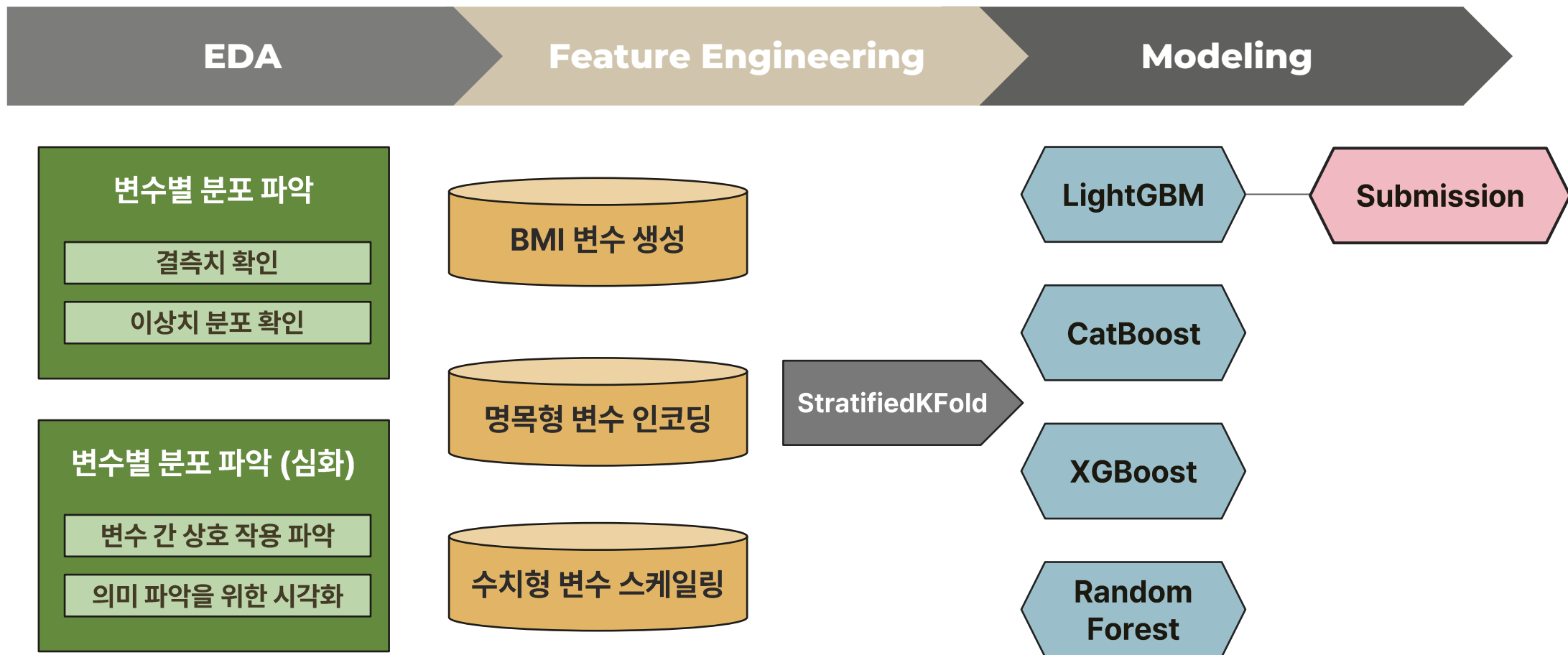
정확도(Accuracy)

- 분류 문제에서 가장 기본적이고 직관적인 성능 평가 지표 중 하나
- 정확도 (*Accuracy*) = $\frac{\text{올바르게 예측된 케이스의 수}}{\text{전체 케이스의 수}} = \frac{TP+TN}{TP+TN+FP+FN}$



		실제 정답	
		True	False
분류 결과	True	TP (True Positive)	FP (False Positive)
	False	FN (False Negative)	TN (True Negative)

3. Modeling Process



Part 2

E.D.A



1. Information about Features in the data

수치형 변수

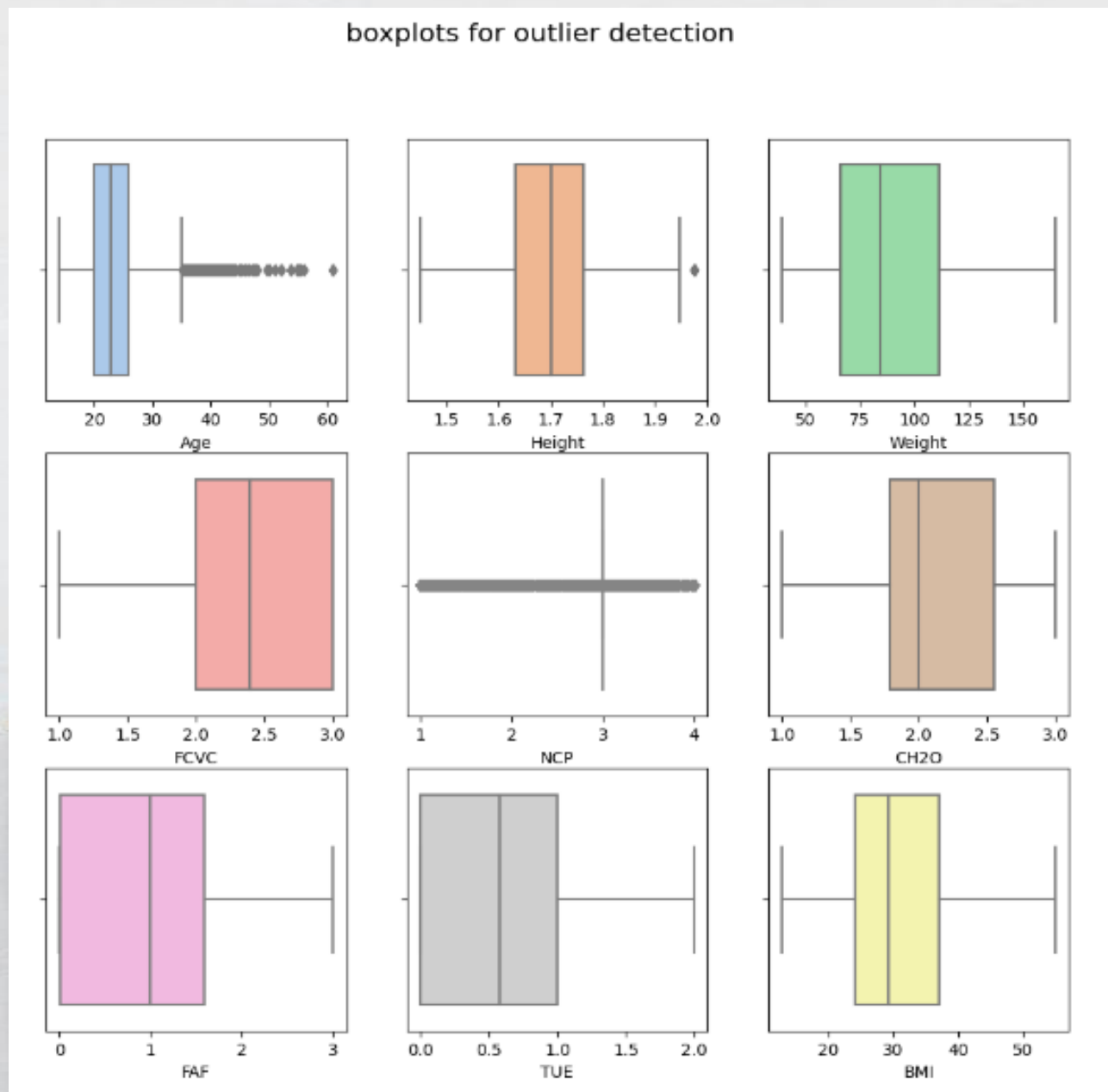
명목형 변수

변수명	내용	변수명	내용
id	인덱스	Gender	성별
Age	나이	Family_history_with_overweight	가족 중 비만이 있는지 여부
Height	키	FAVC	고칼로리 음식 선호 여부
Weight	몸무게	CAEC	식사 사이에 음식을 섭취하는 습관
FCVC	하루에 야채를 섭취하는 빈도	SMOKE	흡연 여부
NCP	하루에 주요 식사를 하는 횟수	SCC	칼로리 섭취량 모니터링 여부
CH2O	하루에 물을 마시는 양	CALC	알코올 섭취 빈도
FAF	일주일에 신체 활동을 하는 빈도	MTRANS	주요 교통 수단
TUE	기술 장치 사용 시간	NObeyesdad	비만 수준, 타겟 변수

2. EDA

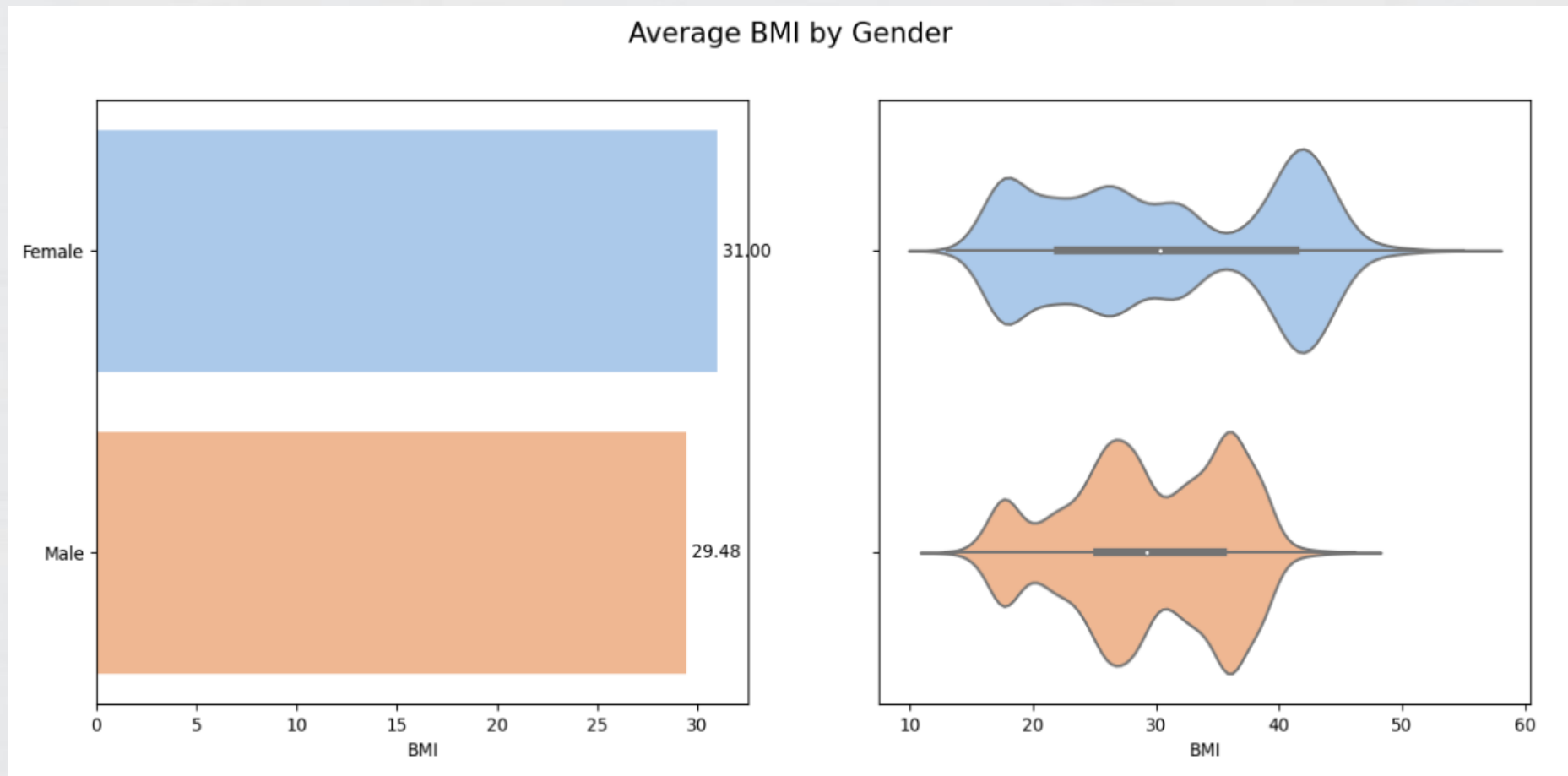
이상치 식별을 위한 박스 플롯 시각화

- 'Age'에서 많은 이상치 발견
- 그러나 최솟값은 14세, 최댓값은 61세로
데이터 설명 내의 범위에 존재 -> 정상적인 값으로 간주



Part 2

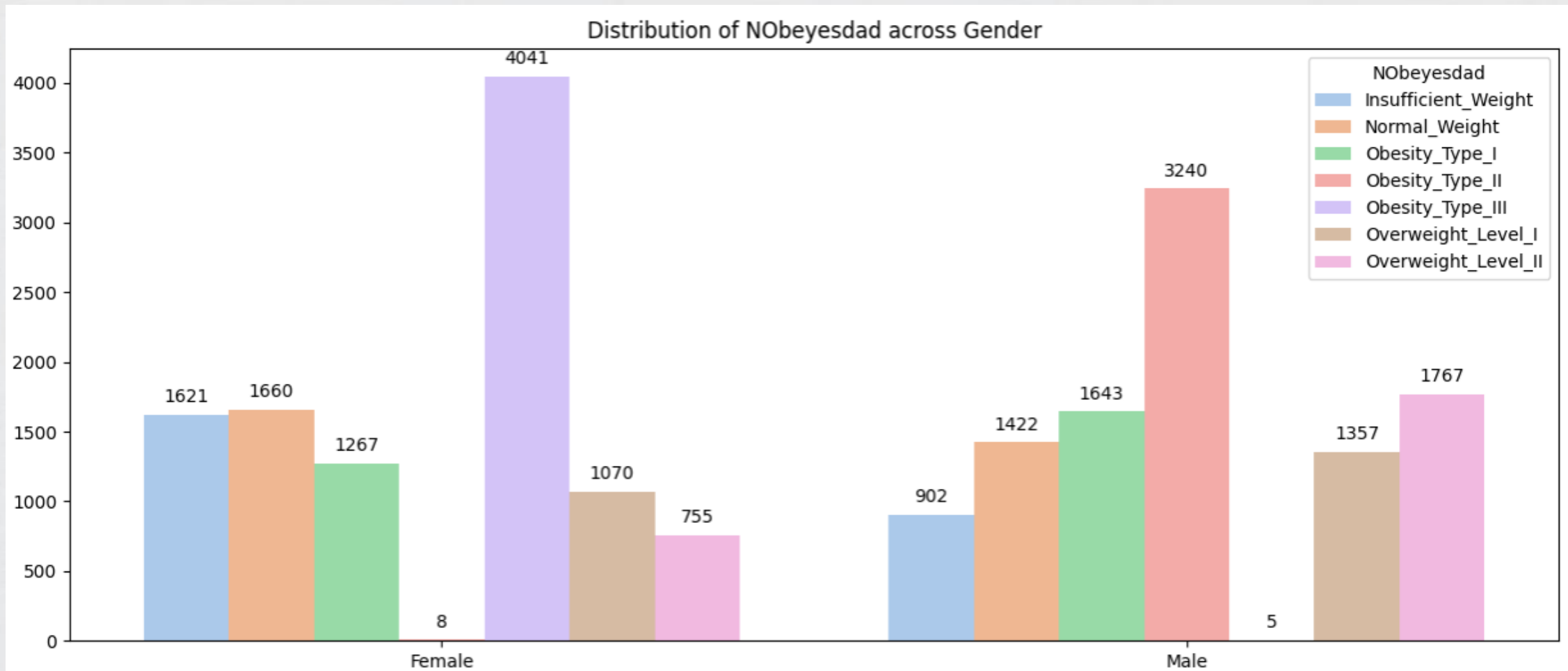
2. EDA



성별에 따른 평균 BMI 시각화

여성 그룹에서 BMI가 다소 높게 나타남

2. EDA



성별에 따른 Nobeyesdad 분포 시각화

- Obesity_Type_III: 주로 여성에게서 관찰됨
- Obesity_Type_II: 주로 남성에게서 관찰됨

Part 3

MODEL EVALUATION



Part 3

Key algorithms for machine learning

RandomForest Classifier

001

XGBoost

002

CatBoost

003

LightGBM

004

1. Random Forest

StandardScaler
OneHotEncoder

BMI

RandomForestClassifier
GradientBoostingClassifier

제목	모델 알고리즘	데이터 재가공	feature engineering	교차 검증	하이퍼 파라미터 튜닝	정확도
RandomForest 1	RandomForest Classifier	StandardScaler : 수치형 데이터 표준화, OneHotEncoder : 범주형 데이터	BMI	StratifiedKFold(n_splits=5, shuffle=True, random_state=42) 0.898640994184376		0.898
RandomForest 1-1	RandomForest Classifier	StandardScaler : 수치형 데이터 표준화, OneHotEncoder : 범주형 데이터	BMI	StratifiedKFold(n_splits=5, shuffle=True, random_state=42) 0.9009052302553556	param_dist ={'n_estimators': 120, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_depth': None}	0.902
RandomForest 1-2. ensemble	RandomForest Classifier, GradientBoostingClassifier	StandardScaler : 수치형 데이터 표준화, OneHotEncoder : 범주형 데이터	BMI		RandomForestClassifier : param_dist ={'n_estimators': 120, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_depth': None}, GradientBoostingClassifier : param_dist ={'n_estimators': 140, min_samples_split : 9, min_samples_leaf :2, max_depth :5}	0.905

2. XGBoost

StandardScaler
OneHotEncoder

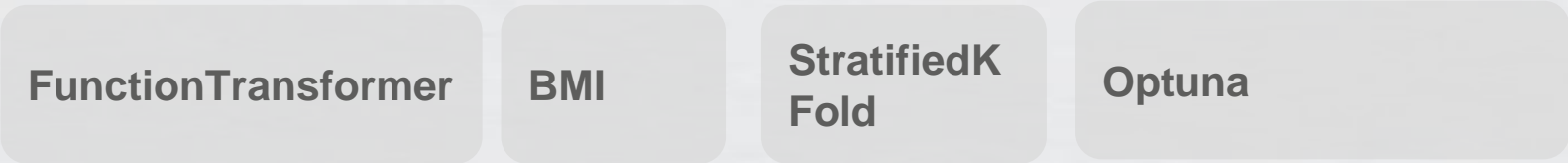
BMI

StratifiedK
Fold

Optuna

제목	모델 알고리즘	데이터 재가공	feature engineering	교차 검증	하이퍼 파라미터 튜닝	정확도
XGBoost 1	XGBoost	StandardScaler: 수치형 데이터 스케일링 OneHotEncoder: 범주형 데이터 인코딩	BMI, 연령대(10단위)	0.904552878	{'classifier__learning_rate': 0.1, 'classifier__max_depth': 5, 'classifier__n_estimators': 200}	0.906069
XGBoost 2	XGBoost	StandardScaler: 수치형 데이터 스케일링 OneHotEncoder: 범주형 데이터 인코딩	BMI, 연령대(10단위)	StratifiedKFold(n_splits=5, shuffle=True, random_state=42) 0.9035891316836775	classifier__n_estimators': randint(100, 1000), 'classifier__learning_rate': uniform(0.01, 0.6), 'classifier__max_depth': randint(3, 10), 'classifier__colsample_bytree': uniform(0.5, 0.5), 'classifier__subsample': uniform(0.5, 0.5)	0.908
XGBoost3	XGBoost	StandardScaler: 수치형 데이터 스케일링 OneHotEncoder: 범주형 데이터 인코딩	BMI	StratifiedKFold(n_splits=5, shuffle=True, random_state=42) 0.9081657944146503	pipeline = Pipeline(steps=[('preprocessor', preprocessor), ('classifier', XGBClassifier(subsample=0.7, n_estimators=900, max_depth=4, learning_rate=0.03, colsample_bytree=0.5, use_label_encoder=False, eval_metric='mlogloss'))])	0.91076

3. Catboost



제목	모델 알고리즘	데이터 재가공	feature engineering	교차 검증	하이퍼 파라미터 튜닝	정확도
Catboost Model	Catboost	FunctionTransformer(age_rounder:Age반올림/height_rounder:Height반올림/extract_features:BMI구하기/col_rounder:FCVC,NCP,CH2O,FAF,TUE반올림) / .select_dtypes(include=['int64','float64']).columns.tolist(): 수치형 데이터 인코딩 / .select_dtypes(include=['object']).columns.tolist() & .remove('NObeyesdad'): 범주형 데이터 인코딩	BMI	StratifiedKFold(n_splits=5, shuffle=True, random_state=42)	CB = make_pipeline(CatBoostClassifier(**params, cat_features=categorical_columns)) params = {'learning_rate': 0.13762007048684638, 'depth': 5, 'l2_leaf_reg': 5.285199432056192, 'bagging_temperature': 0.6029582154263095, 'random_seed': RANDOM_SEED, 'verbose': False, 'task_type':"GPU", 'iterations':1000}	0.911

4. LightGBM

LabelEncoder
루트 변환

BMI

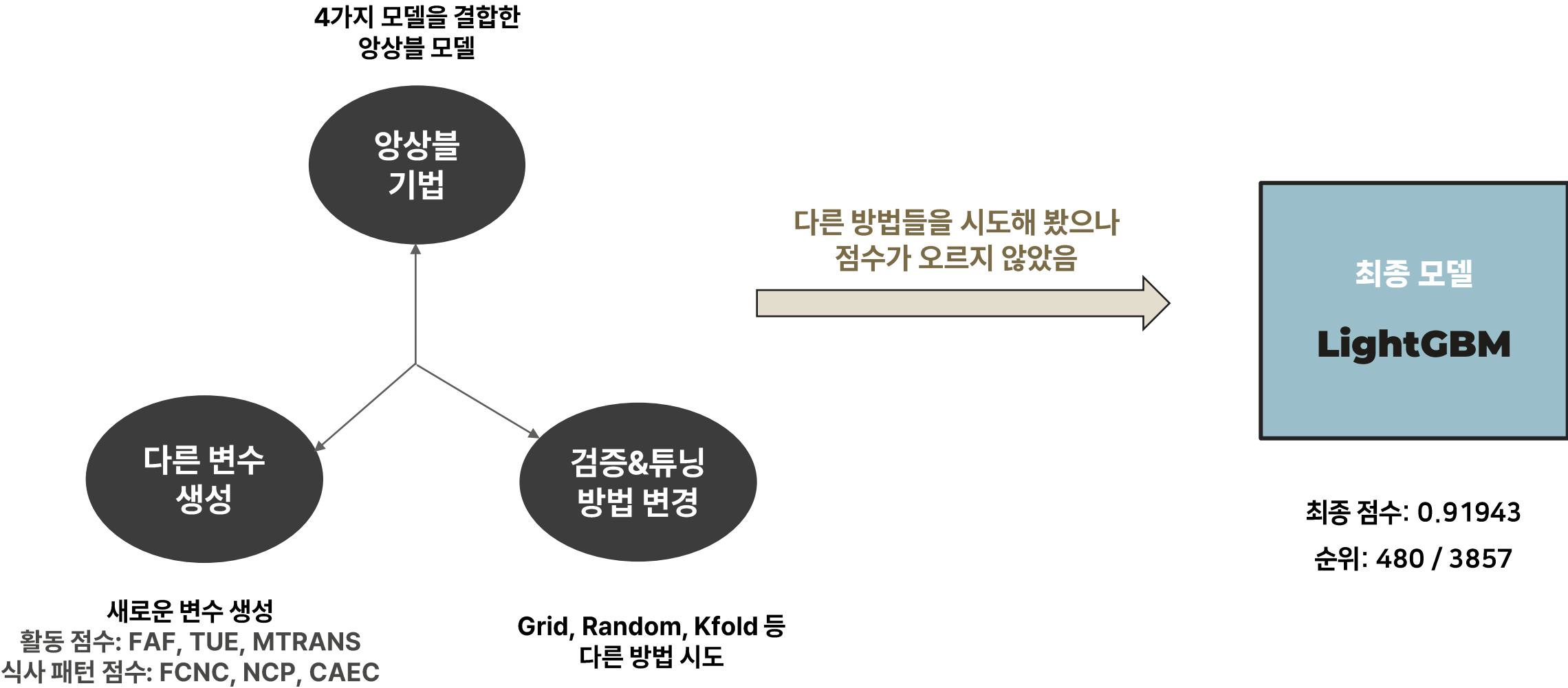
StratifiedK
Fold

Optuna

제목	모델 알고리즘	데이터 재가공	feature engineering	교차 검증	하이퍼 파라미터 튜닝	정확도
LightGBM	LightGBM	결측치 제거, 중복값 제거, LabelEncoder() 사용 scale_cols = ['Age','Height', 'Weight','FCVC','NCP','CH2O',' FAF','TUE']for c in scale_cols: X_train[c] = X_train[c].pow(0.5) X_test[c] = X_test[c].pow(0.5)	BMI	StratifiedKFold(n_spl its=5,random_state= 4,shuffle=True)	{'objective': 'multiclass', 'metric': 'multi_logloss', 'verbosity': -1, 'boosting_type': 'gbdt', 'random_state': 42, 'num_class': 7, 'learning_rate': 0.03096221154683276, 'n_estimators': 500, 'lambda_l1': 0.009667446568254372, 'lambda_l2': 0.040186414373018, 'max_depth': 10, 'colsample_bytree': 0.4097712934687264, 'subsample': 0.9535797422450176, 'min child samples': 26}	0.91943

순위: 480/ 3578

5. Final Model Selction



Conclusion and Insights Summary

데이터 분석

- 정형 데이터이고, 결측치와 중복 값이 없어 분석이 용이
- 비흡연자 분포 과다
- 가상 데이터 특성상 식사 횟수 및 야채 섭취 빈도 등이 소수점으로 나타나 과적합 주의 필요

Feature Engineering

- BMI 생성: 예측에 도움을 준 주요 변수이고, 모델 별 변수 중요도를 시각화 했을 때, Weight 다음으로 변수 중요도 높음
- 명목형 변수 라벨 인코딩
- 수치형 변수 루트 변환: 데이터 분포를 정규 분포에 가깝게 조정, 극단값 영향 감소
- BMI 말고 도움이 되는 다른 변수도 있지 않았을까 하는 아쉬움

모델

- LightGBM: 대 부 분 의 Gradient Boosting 모델과 달리, leaf-wise 성장 전략을 사용해 과적합 위험이 줄어 높은 정확도가 나온게 아닐까 추측
- 개선 가능성: 다른 앙상블 조합 및 스택킹 방법 등을 통한 성능 개선 가능성 존재

Thank you!