


# Kaggle 다중 분류 모델 대회

1조 - 김영기, 김영환, 김진아, 나한울

# 1. 대회 소개


## ◆ Overview of Competition

 KAGGLE · PLAYGROUND PREDICTION COMPETITION · 3 DAYS TO GO

[Submit Prediction](#) ...

## Multi-Class Prediction of Obesity Risk

Playground Series - Season 4, Episode 2




[Overview](#) [Data](#) [Code](#) [Models](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [Submissions](#)

### Overview

**Welcome to the 2024 Kaggle Playground Series!** We plan to continue in the spirit of previous playgrounds, providing interesting and approachable datasets for our community to practice their machine learning skills, and anticipate a competition each month.

**Your Goal:** The goal of this competition is to use various factors to predict obesity risk in individuals, which is related to cardiovascular disease. Good luck!

**Competition Host**

Kaggle 

**Prizes & Awards**

Swag  
Does not award Points or Medals

**Participation**

# 1. 대회 소개

## ◆ 데이터셋 소개

### **the Obesity or CVD risk dataset**

#### **Columns Description**

- id - unique identifier
- Gender - Gender organized as string
- Age - Continuous form of age
- Height - height in meters
- Weight - weight in kg
- family\_history\_with\_overweight
- FAVC - Frequent consumption of high caloric food
- FCVC - Frequency of consumption of vegetables
- NCP - Number of main meals
- CAEC - Consumption of food between meals
- SMOKE - Smoking status
- CH2O - Consumption of water daily
- SCC - Calories consumption monitoring
- FAF - Physical activity frequency
- TUE - Time using technology devices
- CALC - Consumption of alcohol
- MTRANS - Transportation used
- NObeyesdad - Obesity(target)

# 1. 대회 소개

## ◆ 평가지표



머신러닝에는 여러가지의 평가지표가 존재한다.

- 정확도(Accuracy score)
- 정밀도(Precision)
- 재현도(Recall)
- F1 score
- ROC, AUC

우리가 요번의 Kaggle 대회에서 주로 사용한 평가지표는 accuracy score, 정확도를 평가지표로 사용하였다.

# 1. 대회 소개


- 나한울 : LGBM을 RandomSearchCV를 통해서 파라미터를 정의 후, 모델링 진행

	<u>EDA with plotly+ML : Obesity Risk - Version 70</u> Complete · 5d ago	0.90606	0.91907
	<u>EDA with plotly+ML : Obesity Risk - Version 65</u> Complete · 5d ago	0.90724	0.91546

- 김진아 : 종속변수 레이블 인코딩 후, LGBM을 이용하여 모델링하였으며, RandomSearchCV를 사용하여 튜닝 및 교차검증 진행

	<u>[Baseline] Scikit-Learn Pipeline - Version 16</u> Complete · 5d ago	0.90128	0.90715
--	---	---------	---------

- 김영기 : XGBoost와 RandomForest를 soft voting을 이용해서 모델링을 진행  
하이퍼 파라미터는 RandomSearchCV를 통해서 튜닝과 교차검증을 진행

	<u>soft_voting - Version 3</u> Complete · 6d ago · 랜덤포레스트와 xgboost를 soft voting을 이용해서 모델링을함 인코딩과 스케일링을 진행했고 randomsearchcv를 통해서...	0.87897	0.88403
--	---	---------	---------

# 1. 대회 소개

## 최종 모델 선정



EDA with plotly+ML : Obesity Risk - Version 65

Complete · 5d ago

0.90724

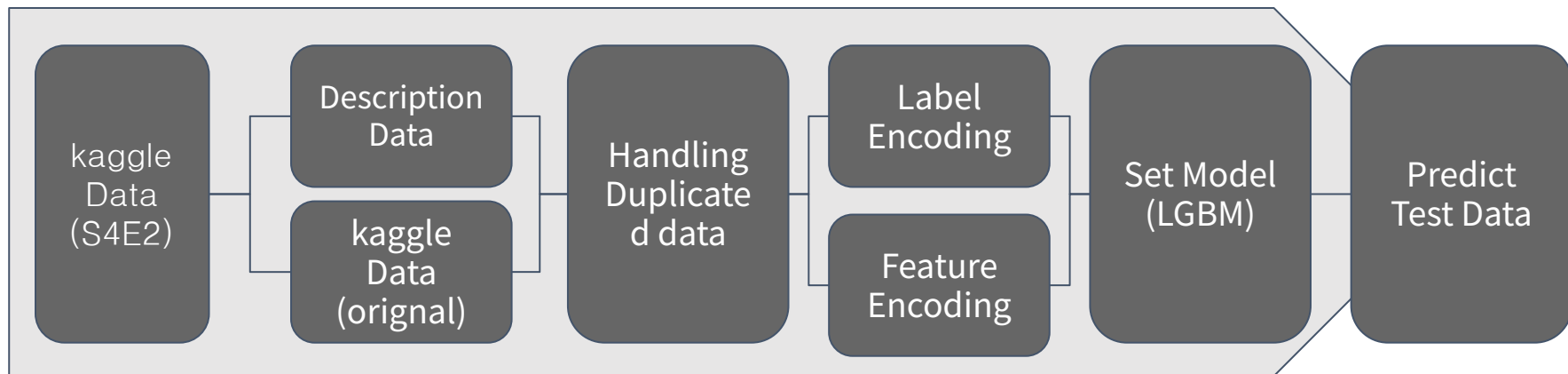
0.91546

[Kaggle Link](#)

- 위의 결과를 봤을 때 LGBM 단일 모델의 정확도가 가장 높았습니다.
- 그렇기에 LGBM 단일 모델을 1조의 대표 모델로 설정하고 진행을 하였습니다.
- PPT에 정리된 모델과 코드는 위의 링크를 첨부하였습니다.

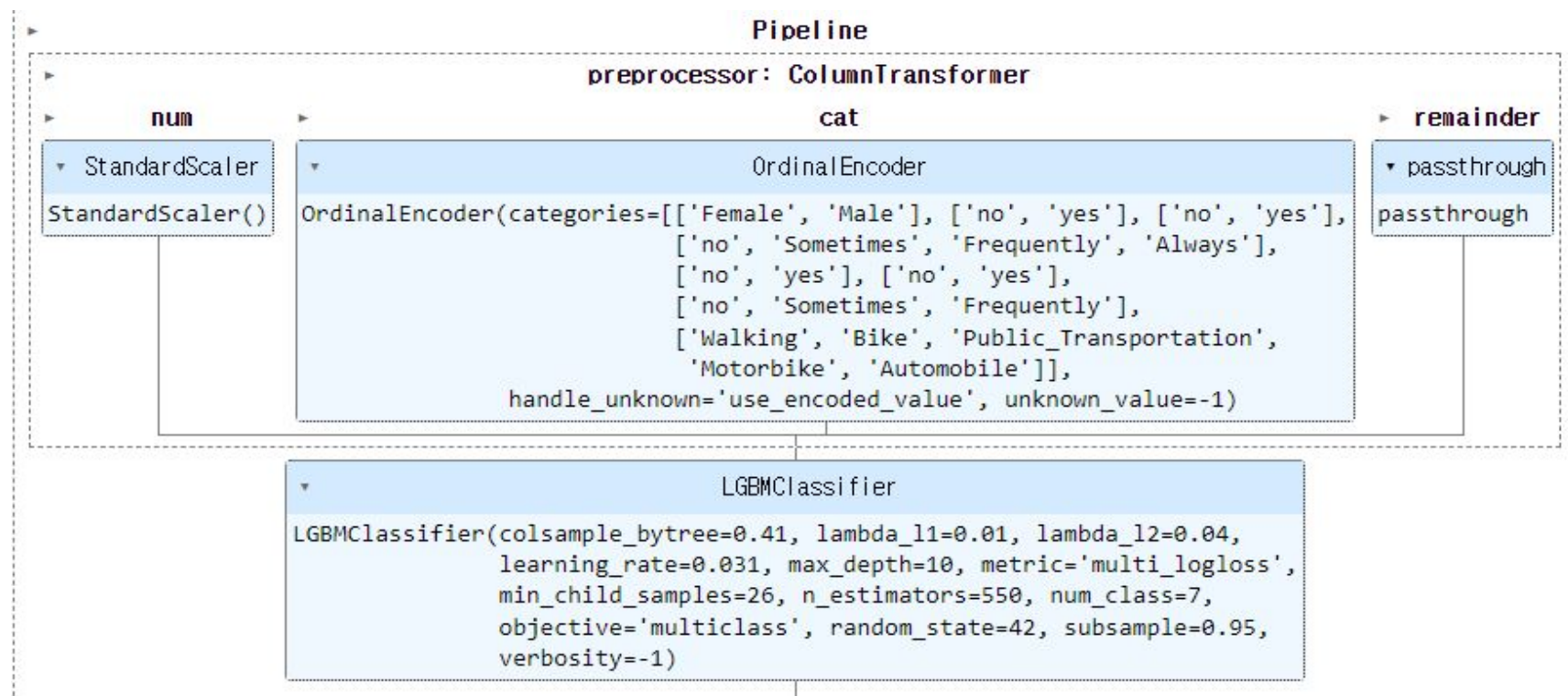
# 1. 대회 소개

- ◆ 모델링 프로세스 / (Architecture) 스케일링 , 데이터 인코딩 코드 설명 (나한울)



# 1. 대회 소개

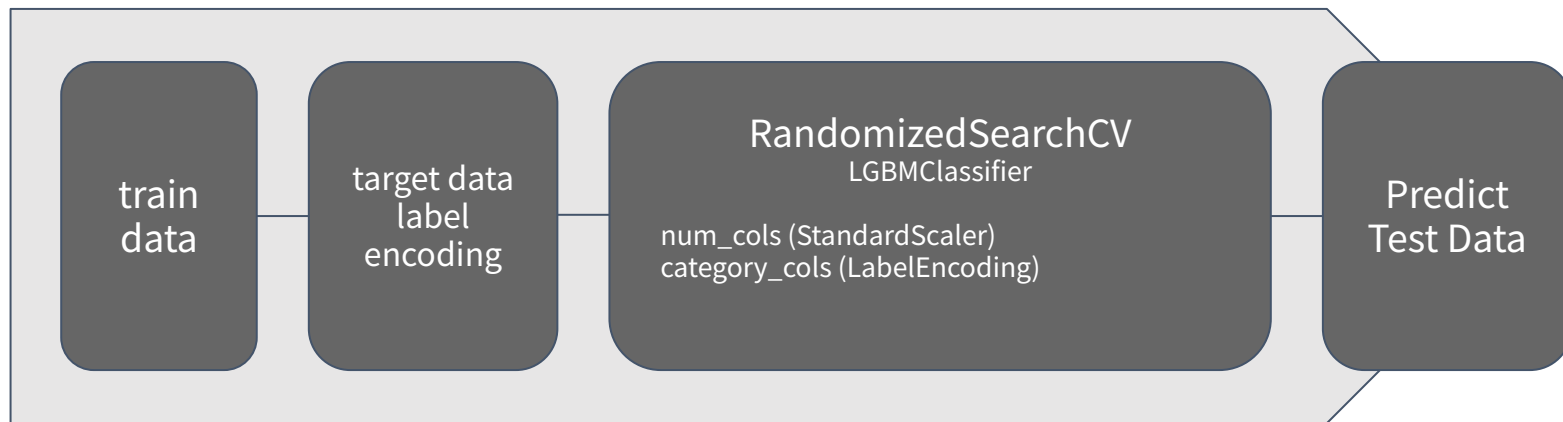
## ◆ 모델링 프로세스 / (Architecture)





# 1. 대회 소개

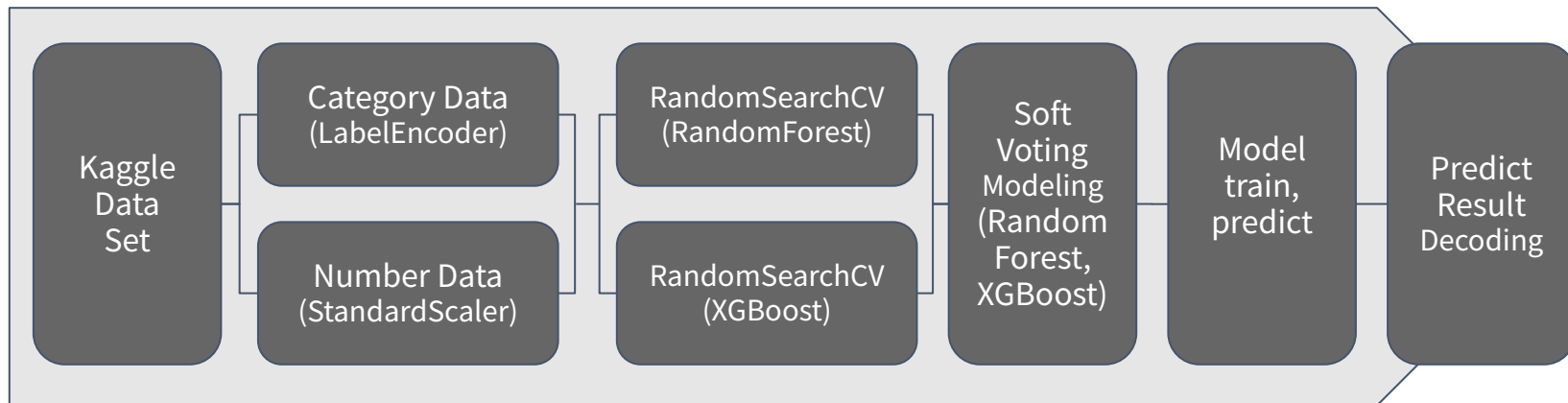
- ◆ 모델링 프로세스 / (Architecture) 스케일링 , 데이터 인코딩 코드 설명 (김진아)



[김진아 Kaggle Link](#)

# 1. 대회 소개

- ◆ 모델링 프로세스 / (Architecture) 스케일링 , 데이터 인코딩 코드 설명 (김영기)



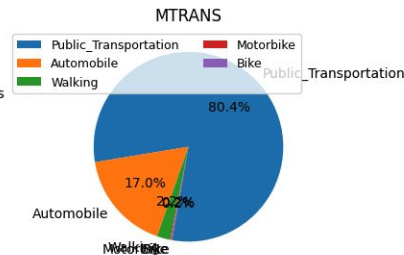
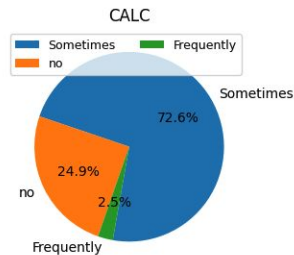
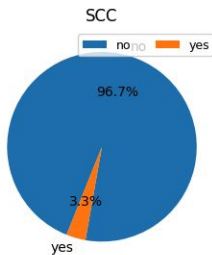
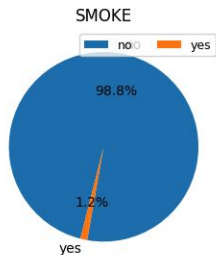
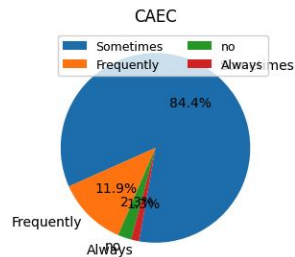
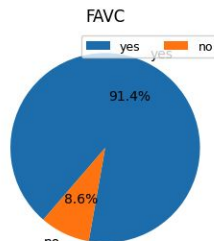
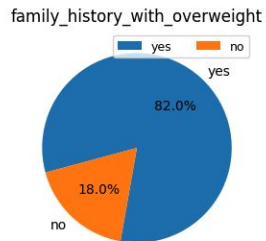
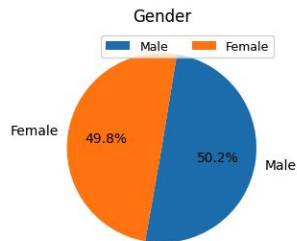
[김영기 Kaggle Link](#)

## 2. 탐색적 자료 분석

### ◆ 주요 시각화 및 통계 분석 보고(카테고리형 피쳐 데이터)

- Gender를 제외한 대부분의 변수들의 값이 불균형하여 학습시 Over Sampling 고려
- 모델링을 할 때, 불균형이 심한 독립변수를 피쳐데이터에서 빼고 학습을 해보았지만 예측정확도에 있어서 큰 변화가 없었다.

Category Feature Data

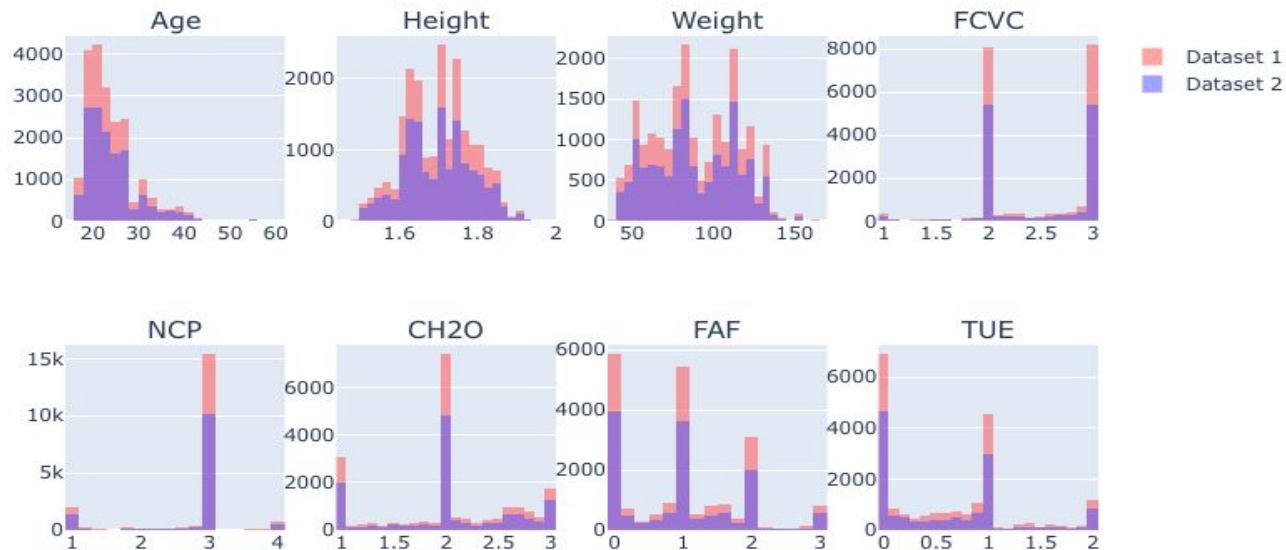


## 2. 탐색적 자료 분석

### ◆ 주요 시각화 및 통계 분석 보고(연속형 피쳐 데이터)

- 대부분 정규분포를 판단하기 어렵지만 어느정도 일정한 형태가 존재함

Multiple Histograms for Different Datasets



### 3. 머신러닝 주요 알고리즘 소개

#### ◆ LGBM(Light Gradient Boosting Machine)

MS가 개발한 고성능 그래디언트 부스팅 프레임워크. 결정 트리기반의 학습 알고리즘이며 대규모 데이터셋을 처리할 때 빠른 학습 속도와 낮은 메모리 사용량을 제공합니다.

#### ◆ XGB(eXtream Gradient Boosting)

양상블 모델의 일종으로 GBM의 기반을 둔 모델입니다. GBM 모델은 부스팅을 기반으로 하는 모델이며, 이전의 학습에서 오답에 대한 가중치를 부여하여 다음 학습시에 학습성능을 높이는 방식입니다. 모든 모델이 연속적으로 학습하며, 서로 의존하는 것을 볼 수 있습니다. 그렇기에 GBM은 학습시간이 길다는 단점이 있는데, 이러한 문제를 고안해서 나온 방법이 XGB입니다.

## 4. 모델 평가

### ◆ 최종모델 선정 과정 (시나리오별 / 모델별)

최종모델  
점수: 0.91907/0.90724  
랭킹: 508/788

Model	Sampling(stratify option)	Hyper Params	Ensemble	Feature Engineering	N Scaling	C Scaling	L Scaling	Accuracy (val)	Accuracy (test)
RFC	X	X	X	X	Standard	OneHot	X	0.886	0.887
RFC	O	X	X	X	Standard	OneHot	X	0.894	0.89
RFC	O	Random	X	X	Standard	OneHot	X	0.894	0.883
RFC,GB,SVM	O	Grid	Voting	X	Standard	OneHot	X	0.899	0.901
XGB	O	Grid	X	Add Data+2	Standard	OneHot	X	0.907	0.907
XGB,RFC,SVM	O	Grid	Voting	Add Data+2	Standard	OneHot	X	0.91	0.9057
LGBM	O	Random(5)	X	Add Data+2	Standard	Ordinal	Hard Coding	0.926	0.908
LGBM	O	Random(10)	X	Add Data+2	Standard	Ordinal	Label(manual )	0.923	0.9122
LGBM	O	Random(10)	X	Add Data+2	Standard	Ordinal	Label(auto)	0.926	0.9151
LGBM	O,SMOTE	Random(10)	X	Add Data+2 Trans Feature	Standard	Ordinal	Label(auto)	0.9225	Overfitting
LGBM	O,SMOTE	Random(10)	X	Add Data+2 Trans Feature	Standard	Ordinal	Label(auto)	0.926	Overfitting
LGBM	O	Random(13)	X	Add Data+2	Standard	Ordinal	Label(auto)	0.926	0.91907

## 5. 결론 및 인사이트 요약

### ◆ 데이터 분석 관련

카테고리형 칼럼들은 Gender를 제외하고 대부분의 값들이 불균형함  
숫자형 칼럼들은 대부분 정규분포를 판단하기 어렵지만 어느정도 일정한 형태가 존재  
시각화나 머신러닝 모델을 만드는데 불균형한 피쳐데이터들의 유무가 모델 정확도의 영향을 끼치지 않음.

### ◆ Feature Engineering 관련

BMI, Age\_Group, MTRANS 단순화, 식습관 점수 등의 파생변수 또는 유니크 값 단순화를 진행하였으나,  
시각화 부분에서도 모델 학습부분에서도 유의미한 결과를 보이지 않음.  
수치형 변수는 Standard로 고정하였으며, 오브젝트 형은 OneHot에서 Ordinal로 바꿨는데, 이유는 학습시에  
순서, 규칙을 주고자함.

### ◆ 모델 관련 (하이퍼파라미터 튜닝 포함)

모델 관련하여 다양한 모델들을 앙상블 한 것보다 LGBM모델을 단일로 사용하였을 때의 정확도가 더 높았는데,  
하이퍼파라미터 튜닝의 문제였을 가능성이 높아보임.

## 6. 이번 캐글대회에서 진행 함으로써 배운 점

**김영기 :** 이번 캐글 대회를 진행하면서 이전에 머신러닝 코드가 이해가 안되었지만 코드에 많이 익숙해졌다. 머신러닝 코드를 짜면서 하이퍼 파라미터를 수없이 바꿔보고 중요하지 않다고 생각하는 피쳐 데이터를 빼보고 새로 추가해 보는 시도를 많이 해봤지만 정확도가 잘 올라가지 않는 에러를 겪었다. 이러한 에러를 해결 하기 위해서 계속적으로 하이퍼 파라미터를 바꿔보고 다른 모델로도 실행을 해보았지만 쉽지 않았다. 이 점으로 다음에 다시 이러한 상황에 놓인다면 이번 대회를 하면서 간과했던 피쳐 데이터들의 상관관계와 통계분석을 신중하게 진행하고, 더 중요한 피쳐 데이터를 만들고 재가공 하는 데이터 재가공 과정에 더욱더 신경을 많이 써야 겠다는 것을 배웠다.

**김진아 :** 이번 대회에는 파라미터 조정과 모델에 집중하여 대회를 진행했었다. 모델의 파라미터를 바꿨을 때에 성과의 차이는 미미했지만 데이터를 가공 후 성과가 조금씩 개선되었다. 이러한 과정을 통해 모델링 전 EDA와 데이터 전처리 과정의 중요성을 깨달았고, 지금까지 배웠던 것들을 잘 정리하여 다음에는 EDA와 데이터 전처리 과정에 조금 더 집중해야겠다고 느꼈다.

**나한울 :** 이제까지 강의에 사용한 데이터, 이미 가공된 데이터들로만 머신러닝을 진행을 해봤어서 모델 구축에도 큰 어려움 없이 할 수 있을 거 같았고, 정확도도 높게 뽑을 수 있을 줄 알았으나 생각보다는 어려움을 겪었음. 이후 하이퍼파라미터 조정과 데이터 인코딩에 대해 더 신경쓰면서 성능은 개선할 수 있었으나, 여전히 0.92를 달성하지는 못해 아쉬움이 존재함.



감사합니다.