

Modeling Team Project Report

: ICR - Identifying Age-Related Conditions

Yeardream 3rd



kaggle

Team NDC

김민주, 박사무엘, 배태양, 이동근, 조인철

0. 목차



1. 프로젝트 소개 및 평가지표

- Target of Project, Evaluation metrics



2. 분석방법론


- EDA & Data Preprocessing, Feature Engineering



3. 이슈사항 및 해결 과정



4. 결과값



5. 결론

- 한계점 및 방향성, 프로젝트 회고

1.1 프로젝트 소개 및 평가지표- Target of Project

ICR - Identifying Age-Related Conditions

: 피실험자를 대상으로 3가지 의학적 상태 중 하나 이상을 가지고 있는지 예측하는 프로젝트

Target of Project

- 1) Public LB와 Log Loss 간 Score 줄이기
- 2) Public LB Score 0.12~0.15 기록하기

1.2 프로젝트 소개 및 평가지표 - Evaluation metrics

$$\text{Log Loss} = \frac{-\frac{1}{N_0} \sum_{i=1}^{N_0} y_{0i} \log p_{0i} - \frac{1}{N_1} \sum_{i=1}^{N_1} y_{1i} \log p_{1i}}{2}$$

참조

N_0, N_1 : 데이터 포인트 수

y_{0i}, y_{1i} : 데이터 포인트의 실제 레이블 값

p_{0i}, p_{1i} : 데이터 포인트에 대한 모델의 예측 확률

Log Loss

: 분류 문제에서 모델 성능 측정을 위해 사용하는 손실 함수

2. 분석방법론

1) Data Information

- Data : $617 * 58$
- Row : 617
- Column : ID, class in 58 columns

2) Vector & Target

- Input Vector : ***Columns*** in Data
- Target Value : ***Class*** (0 or 1 by the three medical conditions)

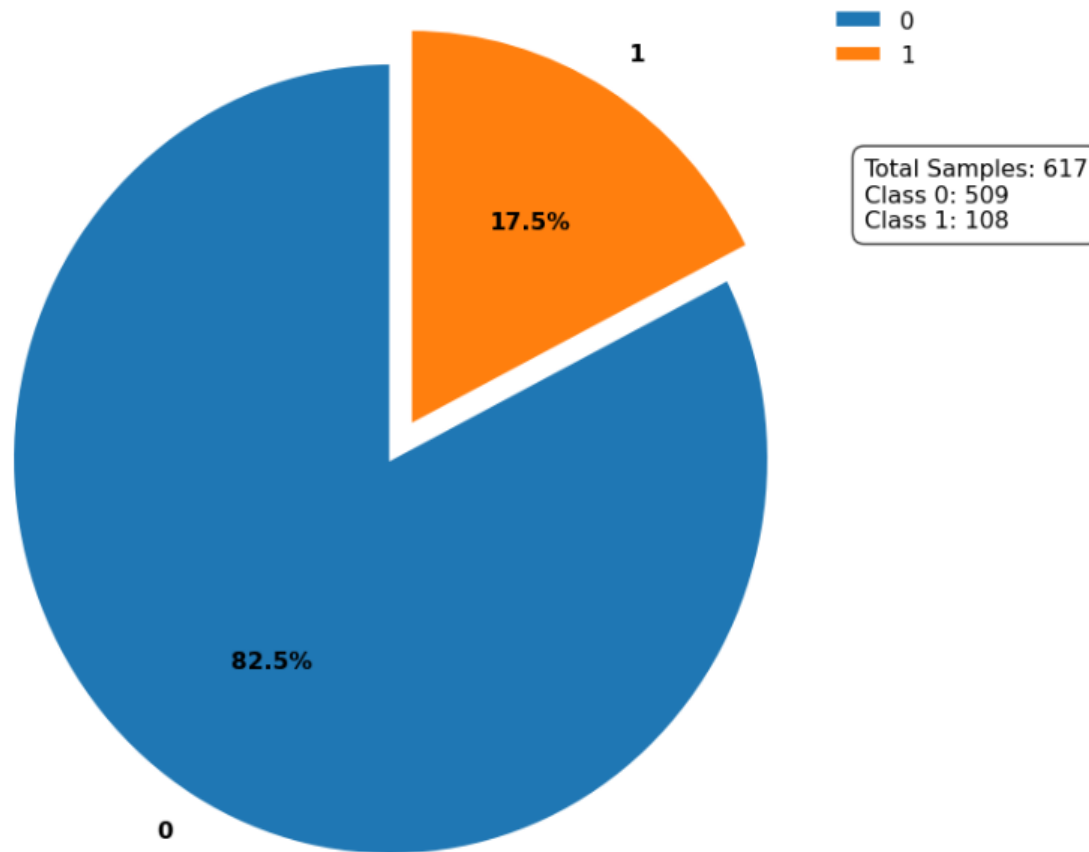
3) Used for learning

- X : 'better' Feature selection(40) by discussion
- y : Class

cf) Greeks data : 과적합 이슈로 우선 제외하고 모델링

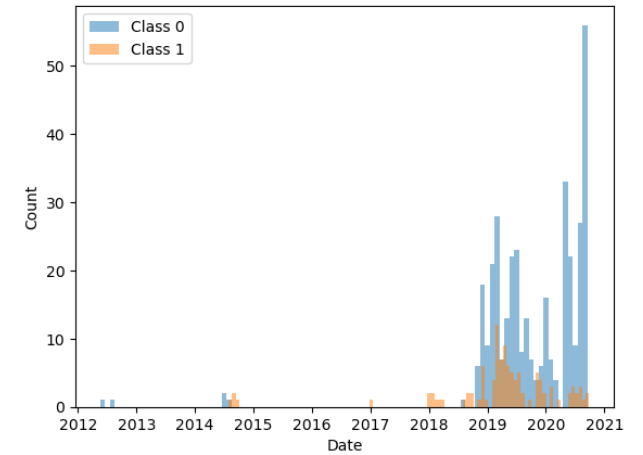
2.1 분석방법론 - EDA(1-1)

Class Distribution

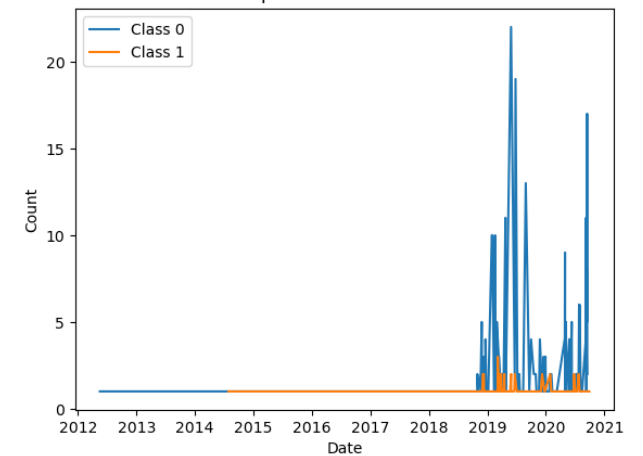


Class 값 차이

Distribution of Dates

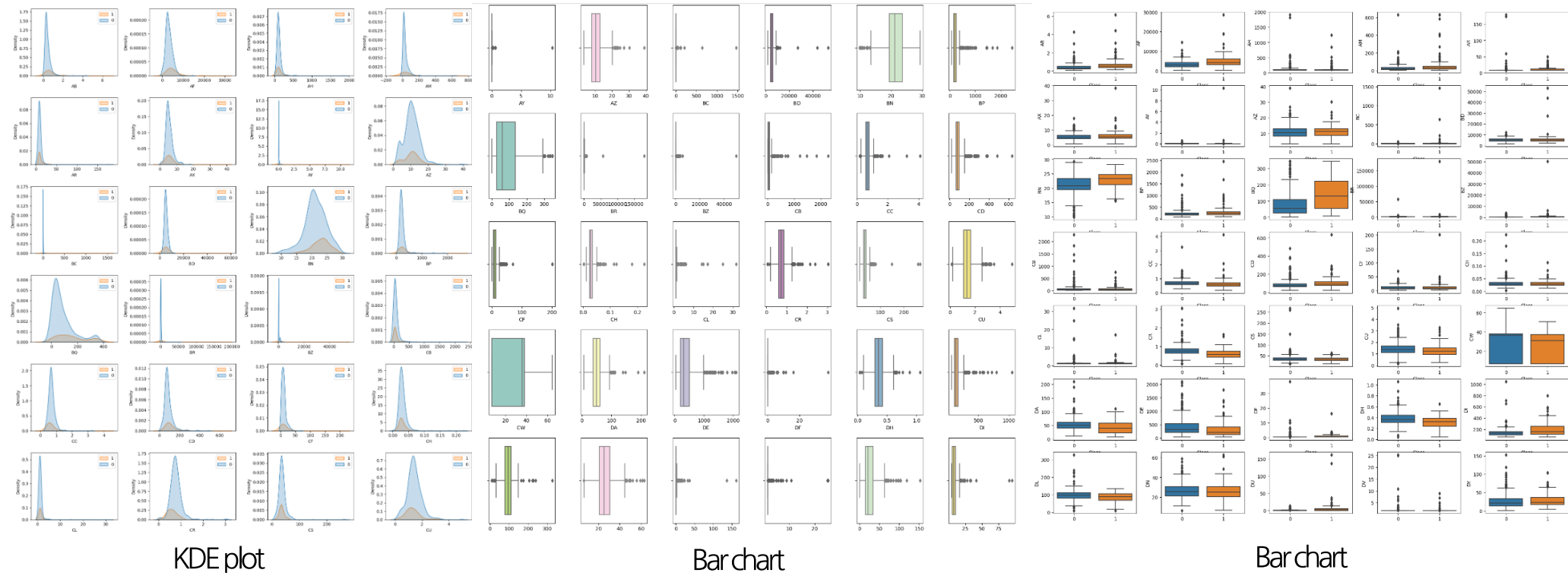


Comparison of Date Counts



연도 별 Class 비교

2.1 분석방법론 - EDA(1-2)

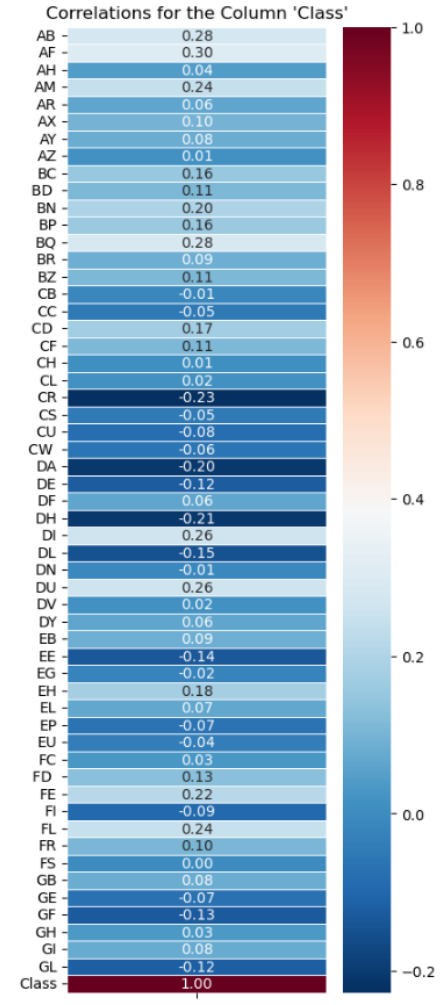
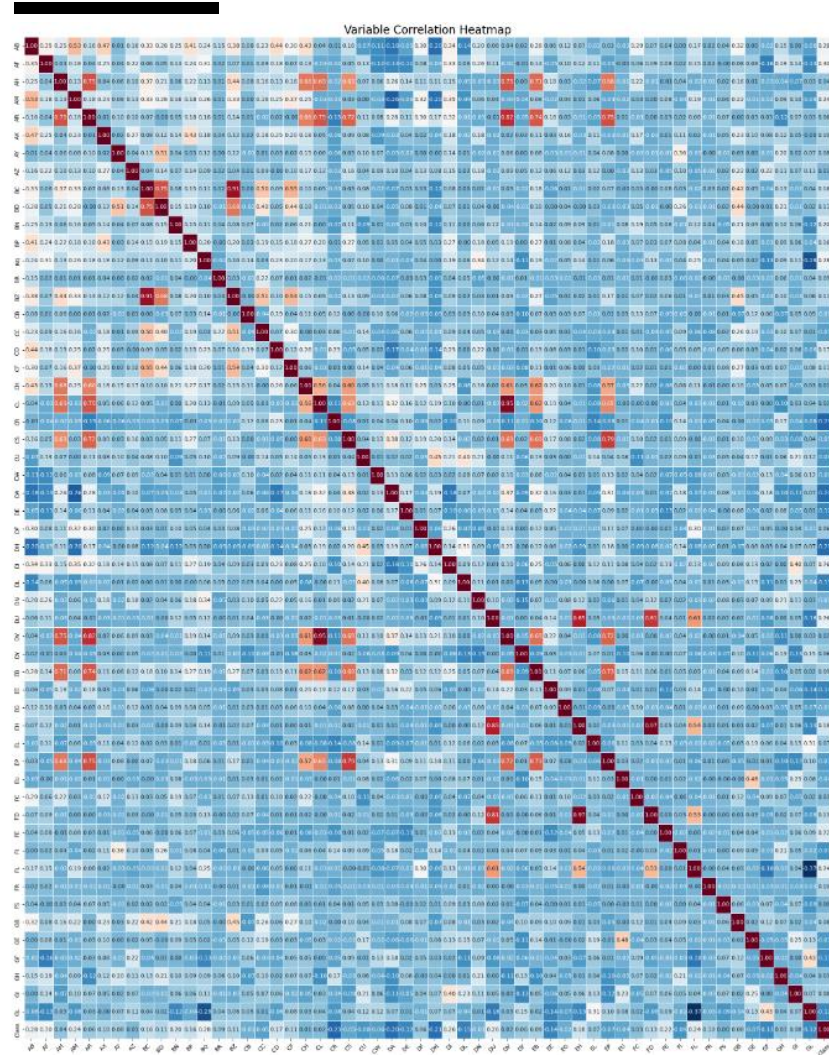


KDE plot

Bar chart

Bar chart
by class

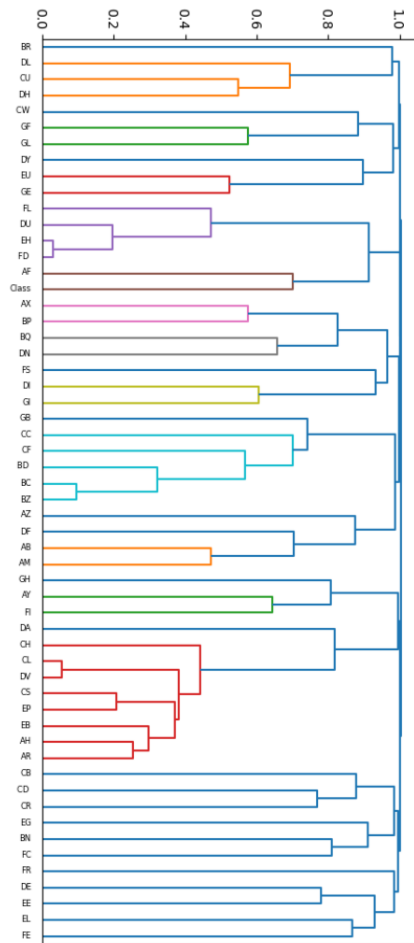
2.1 분석방법론 - EDA(2-1)



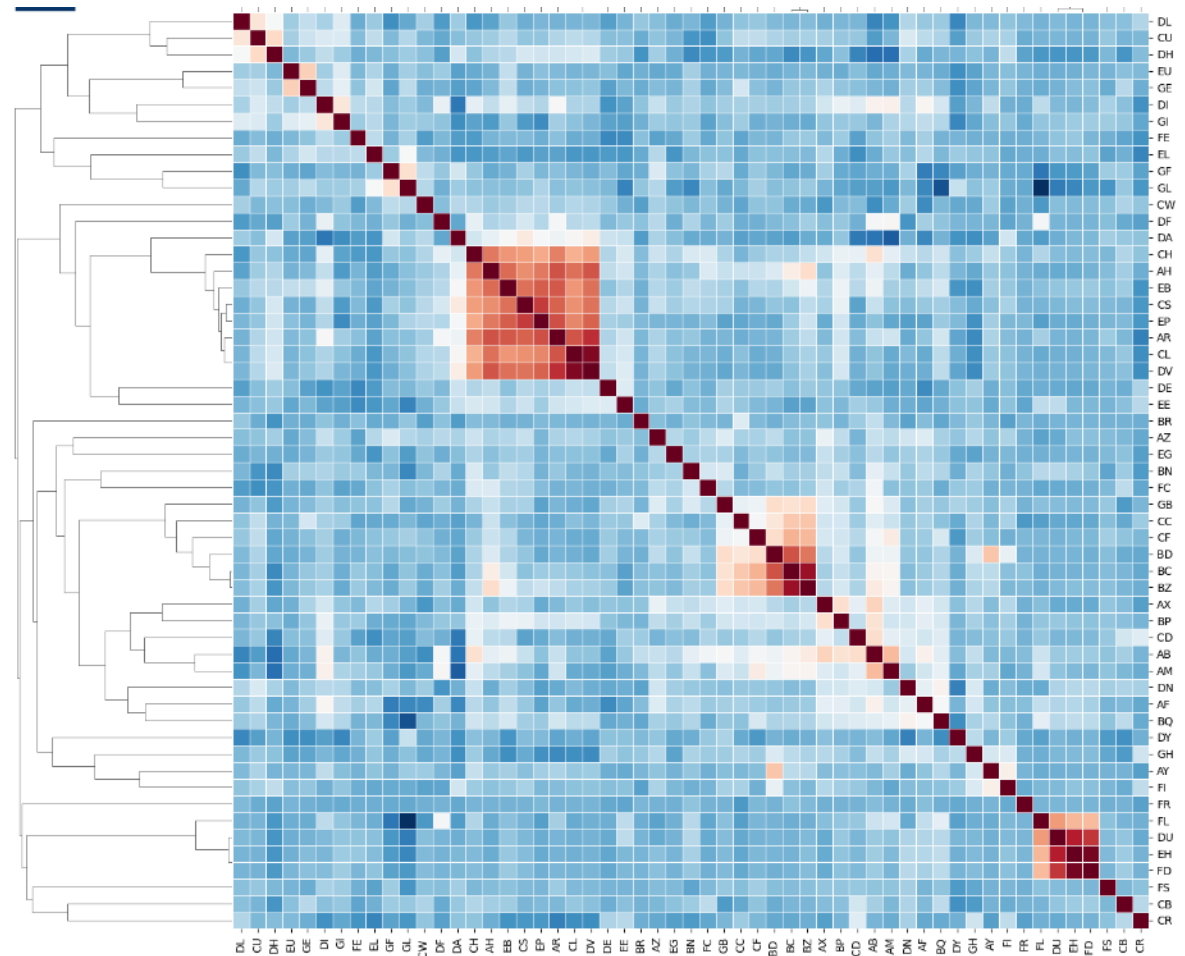
상관관계

Class 상관관계

2.1 분석방법론 - EDA(2-2)



Hierarchical clustering
- 군집화 관계성 파악



상관관계 + 군집화

2.2 분석방법론 - Data Preprocessing, Feature Engineering

- 전처리 (Data Preprocessing)

- 1) LabelEncoder
- 2) KNN- Imputer
- 3) Feature Importance
- 4) VIF (다중공선성)
- 5) Class Imbalance Handling - Over Sampling
- 6) Outlier

- Feature Engineering

- 1) 방향성
 - Better score
 - Do not over-fitting
- 2) 한계
 - blind or unknown data
 - Meta Data → Greeks Data 사용에 어려움

3. 이슈사항 및 해결 과정

0) Debugging

1) Used Model

- XGBOOST
- TabPFN
- RandomForest
- LightGBM
- Catboost

2) Final Hyper-Parameter

- Optuna
- Default Set Value
























3) Performance

Total Log loss : 0.05

Avg Loss : 0.0502
Total logloss : 0.0502

4. 결과값 (현재 기준)

Using Ensemble Model
Public LB Score : 0.18

	SVM+RF+XGB KFold ensemble 0.1 제발 - Version 3 Succeeded · 5h ago · Notebook SVM+RF+XGB KFold ensemble 0.1 제발 Version 3	0.27		SVM+RF+XGB KFold ensemble 0.1 제발 - Version 3 Succeeded · 5h ago · Notebook SVM+RF+XGB KFold ensemble 0.1 제발 Version 3	0.27
	SVM+RF+XGB KFold ensemble 502ebe - Version 1 Succeeded · 2d ago · Notebook SVM+RF+XGB KFold ensemble 502ebe Version 1	0.22		SVM+RF+XGB KFold ensemble 502ebe - Version 1 Succeeded · 2d ago · Notebook SVM+RF+XGB KFold ensemble 502ebe Version 1	0.22
	SVM+RF+XGB KFold ensemble 29975b - Version 2 Notebook Threw Exception · 2d ago · Notebook SVM+RF+XGB KFold ensemble 29975b Version 2			SVM+RF+XGB KFold ensemble 29975b - Version 2 Notebook Threw Exception · 2d ago · Notebook SVM+RF+XGB KFold ensemble 29975b Version 2	
	SVM+RF+XGB KFold ensemble fd2484 - Version 2 Notebook Threw Exception · 3d ago · Notebook SVM+RF+XGB KFold ensemble fd2484 Version 2			SVM+RF+XGB KFold ensemble fd2484 - Version 2 Notebook Threw Exception · 3d ago · Notebook SVM+RF+XGB KFold ensemble fd2484 Version 2	
	Fork of [for Beginner] SVC+RF+LR+XGBwith A 7d519c - Version 3 Succeeded · 3d ago · Notebook Fork of [for Beginner] SVC+RF+LR+XGBwith A 7d519c Version 3			Fork of [for Beginner] SVC+RF+LR+XGBwith A 7d519c - Version 3 Succeeded · 3d ago · Notebook Fork of [for Beginner] SVC+RF+LR+XGBwith A 7d519c Version 3	0.33
	 SVM+RF+XGB KFold ensemble - Version 34 Succeeded · 5h ago · Notebook SVM+RF+XGB KFold ensemble Version 34			0.18	0.27
	SVM+RF+XGB KFold ensemble(mule) - Version 2 Succeeded · 5h ago · Notebook SVM+RF+XGB KFold ensemble(mule) Version 2	0.24		SVM+RF+XGB KFold ensemble 180561 - Version 2 Succeeded · 11h ago · Notebook SVM+RF+XGB KFold ensemble 180561 Version 2	0.26
	SVM+RF+XGB KFold ensemble 6/28 best - Version 4 Succeeded · 1d ago · Notebook SVM+RF+XGB KFold ensemble 6/28 best Version 4	0.29		SVM+RF+XGB KFold ensemble 628 (3) - Version 12 Succeeded · 1d ago · best default of 628	0.29
	ICR PJT RF+XGB+TABPFN(hybrid) - Version 3 Succeeded · 6d ago · Notebook ICR PJT RF+XGB+TABPFN(hybrid) Version 3	0.25		SVM+RF+XGB KFold ensemble 0626 - Version 2 Notebook Threw Exception · 3d ago · Notebook SVM+RF+XGB KFold ensemble 0626 Version 2	
	Fork of [for Beginner] SVM+RF+LR+XGB+NN(hybrid) - Version 3 Notebook Threw Exception · 7d ago · Notebook Fork of [for Beginner] SVM+RF+LR+XGB+NN(hybrid) Version 3			SVC+RF+LR+XGBwith Auto 6.24 w.samuel - Version 2 Succeeded · 5d ago · Notebook SVC+RF+LR+XGBwith Auto 6.24 w.samuel Version 2	0.34
	[for Beginner] SVM+RF+LR+XGB+NN with AutoML - Version 1 Succeeded · 8d ago · Notebook [for Beginner] SVM+RF+LR+XGB+NN with AutoML Version 1	0.21		ICR PJT RF+XGB+TABPFN6.23 - Version 1 Succeeded · 7d ago · Notebook ICR PJT RF+XGB+TABPFN6.23 Version 1	0.22
	ICR PJTsun - Version 2 Notebook Threw Exception · 10d ago · Notebook ICR PJTsun Version 2				

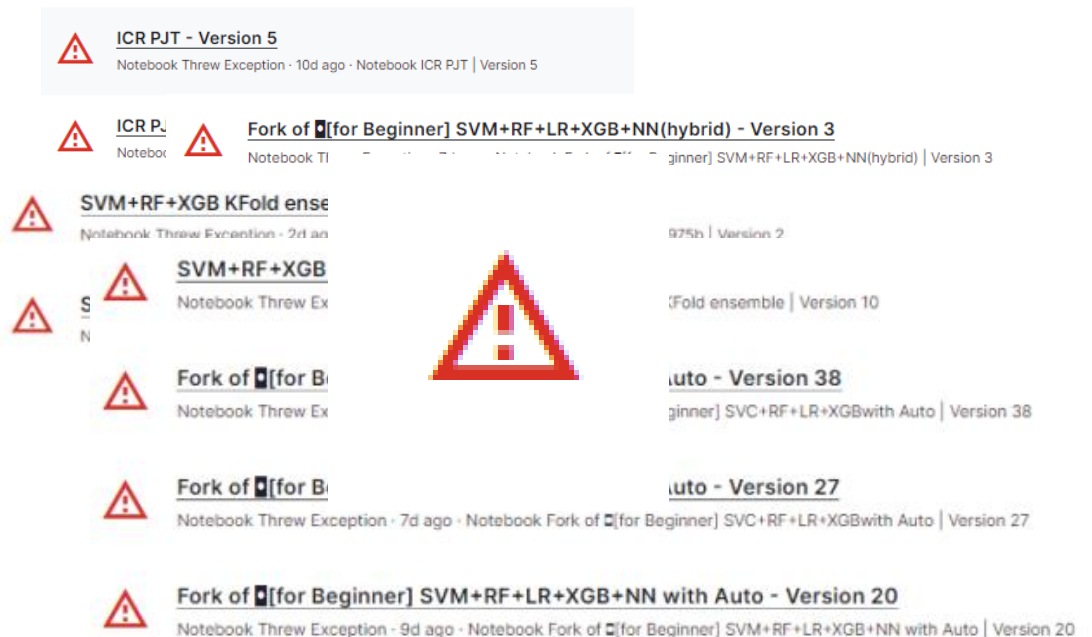
5.1 결론 - 한계점 및 방향성

1) Risk & failed method

- Data Feature
- Submit Problem
- Debugging
- Unused Meta Data (Greeks)
- OOF & LB score Balance

2) 방향성

- Meta Data
- Best Ensemble
- Over sampling
- Discussion
- Epsilon



5.2 결론 - 프로젝트 회고

- 1) **전처리, 디버깅** 과정 중 많은 시행착오 겪고, 많은 시간을 소요하였다. 이에 본격적인 모델링전에 데이터를 탐색하는 과정이 굉장히 기본적인자 중요한 **‘핵심역량’**임을 다시금 확인하는 시간이었다.
- 2) Blind or unknown 데이터를 다루며 앞으로 이러한 데이터를 다룰 수 있겠다 싶어 암담했다. 어려웠지만, **연구자의 직관에 의존한 해석**이 경계해야 할 영역임을 배울 수 있었다.
- 3) 코드 실행하고 Submit 까지 기회가 많지 않아 굉장히 난감했다. 이를 통해 사실상 코드를 작성하고 수정하는 것 뿐만 아니라 온전한 코드실행을 기대하는 인고 통해 **기다림의 미학**을 느꼈다.
- 4) Ensemble 과정 등 더 나은 모델, 더 좋은 성능을 찾기 위해 **최신 논문, 기술들을 숙지할 필요성**이 있었다.
- 5) 자신의 판단을 기반으로 데이터 분석, 파라미터 및 모델 조정 등의 과정에서 **기대치**를 높여 가는 부분이 쉽지 않았다. 데이터를 얼마나 아느냐에서 부터 코드를 세세히 뜯어보는 것 까지 프로젝트 과정에서 내가 할 수 있는 것을 많이 고민하며 배우는 시간이었다.