AI 머신러닝 프로젝트 보고서

효소 안정성 예측

7조
김효진, 나다경, 안이현, 유도현

목 차

1. 서론 ····· **3**

1.1 연구배경 ····· **3**

1.2 목표 ····· **3**

2. 방법론 ····· **4**

2.1 XGBoost ····· **4**

2.2 LightGBM ····· **4**

2.3 Ridge ····· **4**

2.4 Random Forest ····· **4**

2.5 KFold vs GroupKFold ····· **4**

3. 실험 ····· **5**

3.1 실험 설정 ····· **5**

3.1.1 데이터 설명 ····· **5**

3.1.2 EDA 및 전처리 ····· **6**

3.1.3 새로운 훈련 데이터 생성 ····· **7**

3.1.4 데이터 특성 선택 ····· **8**

3.1.5 모델링 ····· **10**

3.2 실험 결과 ····· **11**

3.2.1 Feature Importance ····· **11**

3.2.2 Best Model ····· **11**

4. 결론 ····· **12**

1. 서론

1.1. 연구 배경

다양한 산업에 쓰이는 효소의 열 안정성은 특정 환경에서 안정적이지 않아, 세포에서 생성될 수 있는 단백질의 양을 감소시킨다. 효소는 단백질 중합체이기 때문에, 단백질 안정성을 효율적으로 예측할 수 있다면 효소가 쓰이는 산업의 발전과 비용 절감 효과를 기대할 수 있어 해당 주제를 선정하였다. 최근에 FoldX, Rosetta와 같은 물리학 원리에 기반한 방법이 개발되는 등 눈에 띄게 발전된 방법으로 단백질 안정성이 예측되고 있다. 최근에는 야생형의 변이 패턴과 3차원 구조를 기반으로 돌연변이가 단백질에 미치는 열 안정성 영향을 예측하기 위해 AlphaFold2와 같은 딥러닝 기반 모델이 제안되어 기존에 있던 문제를 어느 정도 해결하고 있다. 하지만, 현재 존재하고 있는 단백질 안정성을 예측하는 머신러닝 알고리즘은 대부분 훈련 데이터의 수가 충분하지 않고, 단백질의 어떤 특성이 안정성과 연관되어 있는지에 대한 연구가 부족하여 과적합 문제가 개선되지 않은 상태이다. 이에 우리는 견고한 알고리즘을 개발하기 위한 '유용한 특성'을 늘리는 것에 중점을 두고, 널리 알려진 머신러닝 기법으로 모델링하는 대신 데이터의 Feature Engineering 과정에 집중하여 단백질의 열안정성에 영향을 미치는 변수를 탐구하기로 하였다.

1.2. 목표

본 대회의 목표는 단백질의 열안정성(tm)을 예측하는 것이다. 안정적인 효소를 찾아야 하는 대회 목적에 따라 열안정성이 상대적으로 더 높은 효소 구조를 발견하는 것이 중요하므로, 특정 구조를 가진 단백질의 열안정성 자체를 예측하는 것이 아닌 순위를 정확히 평가하는 것이 목표이다. 높은 tm 값은 단백질이 상대적으로 더 안정적이라는 것을 의미하므로 더 안정적인 단백질에 높은 순위를 주어 순위에 대한 스피어만 상관계수를 최대한 높이는 것이 본 분석의 최종 목표이다. 따라서 tm의 정확한 값보다는 예측 결과의 상대적인 순서가 더 중요하다.

2. 방법론

2.1 XGBoost

XGBoost는 Extreme Gradient Boosting의 약자로 Gradient Boosting 알고리즘을 병렬 학습이 지원되도록 구현한 Decision Tree 기반 앙상블 모델이다. 여기서 Boosting이란 여러 개의 약한 예측 모형들을 조합해서 사용하는 Ensemble 기법의 하나로, 초반에 생성되는 모델들의 학습 에러에 가중치를 두고, 순차적으로 다음 학습 모델에 반영해 예측 모형의 성능을 끌어올리는 방법이다. XGBoost를 사용한 이유는 병렬로 학습을 하기 때문에 수행시간이 빨라서 모델을 효과적으로 실험해볼 수 있었기 때문이다. 또한 XGBoost는 자체에 포함된 과적합 규제 기능 때문에 상당히 견고하며, 의사결정나무 기반 앙상블 모델 특유의 뛰어난 예측 성능을 발휘할 수 있다.

2.2 LightGBM

LightGBM이란 일반적인 Gradient Boosting처럼 Tree의 균형을 최대한 유지하면서 깊이를 늘려나가지 않고 비대칭적인 Tree 분할을 통해 Tree의 깊이를 손쉽게 늘림으로써 예측 오류를 최소화하는 모델이다. LightGBM을 사용한 이유는 하나의 분할에서 생성되는 두 노드 중 손실이 큰 노드에서만 분할을 이어 나가기 때문에 학습 속도가 빠르고 메모리 사용량이 적어 고성능 하드웨어를 갖추지 못한 환경에서 적절한 머신러닝 모델이 될 것이라 생각했다.

2.3 Ridge

대표적인 정규화 방식 중 하나인 Ridge는 기존의 오차제곱합을 이용하는 회귀에 미분 가능한 penalty term을 추가하여 학습에 불필요한 변수의 가중치를 0에 가깝게 유지하는 방식이다. Ridge를 사용한 이유는 모델의 과적합을 방지하는 동시에, 단백질 시퀀스에서 추출해낸 feature들 사이에 다중공선성 문제가 있어도 효과적으로 예측을 수행하기 때문이다.

2.4 Random Forest

Random Forest는 하나의 데이터 세트에서 많은 의사결정나무를 생성하는 기존의 Bagging에서 각 노드 분할 때의 변수 선택에 제한을 둠으로써 일반화 성능을 끌어올린 모델이다. Random Forest는 좋은 성능을 발휘하면서 과적합은 감소시키고, 변수의 중요성을 파악할 수 있기 때문에 이 연구에서도 해당 방법을 사용했다.

2.5 GroupKFold

Test set에는 구조가 비슷한 단백질들이 존재하기 때문에 우리는 Feature Engineering 과정에서 train set 내의 단백질들을 구조가 비슷한 것들끼리 그룹화하였으며, Cross-validation 과정에서 이를 활용하고자 하였다. 일반 k-Fold CV는 랜덤하게 폴드를 결정하기 때문에 데이터가 그룹화되어 있을 경우 그룹 내 데이터가 k개의 validation set에 여러 번 존재하게 된다. 하지만 GroupKFold는 그룹 내 모든 데이터가 단 하나의 validation set에만 등장하면서 validation set 내 그룹이 최소 개수로 존재하도록 통제되기 때문에 test set과 validation set의 차이가 줄어들어 과적합 문제를 해결할 수 있다.

3. 실험

3.1 실험 설정

3.1.1 데이터 설명

그림 1. Train, Test set

Train set

Train set은 총 5가지 변수로 구성되어 있다. 그 종류로는 실험 단백질의 ID인 'seq_id', 단백질의 1차원적 구조를 나타내는 'protein_sequence', 단백질 안정성이 측정되는 수용액의 산성도를 나타내는 'pH', 단백질이 연구자들에 의해 실험된 출처를 표시한 'data_source', 그리고 우리가 예측할 타깃 변수인 'tm'이 있다. tm은 각 단백질의 구조가 열에너지에 의해 깨어지는 순간의 온도 값(melting temperature)을 의미한다. 단백질 구조의 변화가 열 안정성에 영향을 미친다는 사실이 이미 알려져 있기 때문에, 주어진 데이터 세트에서 가장 중요한 변수는 'protein_sequence'이다. 단백질은 아미노산들의 결합에 의해 생성되기 때문에 해당 변수는 이를 표현하고자 단백질을 구성할 수 있는 20가지 아미노산들을 일렬로 나열한 문자열 형식으로 되어있다. 각 아미노산은 어떤 아미노산과도 결합할 수 있고 단백질 내에서 중복하여 존재할 수 있기 때문에 단백질마다 길이와 구조는 아주 다양하다.

Test set

test set은 한 가지의 야생형 효소(VPVN…VGK)로부터 1개씩의 아미노산만 변형된 2413개의 단일 돌연변이들로 구성되어 있다. 데이터 출처와 pH는 모두 동일하다.

PDB file

단백질 분자의 3차원적 구조를 설명하는 단백질 정보 파일이다.

External Data

우리는 외부 데이터 세트 중에서 아미노산의 분자 무게, 아미노산 잔기 등 단백질을 구성하는 아미노산에 대한 17가지 성질이 담긴 CSV 파일을 불러와 'aa_props' 변수에 저장하여 사용하였다. 또한 본 대회는 train set과 AlphaFold 구조 예측 결과 외에 ESM, EVE, Rosetta 등 단백질 안정성을 예측할 수 있는 공개된 외부 모델의 사용을 허용하고 있지만, 우리는 대회에서 제공한 AlphaFold 외에는 사용하지 않았다. 이러한 모델들의 경우 이미 연구가 완료되어 단백질 열 안정성과 직접적으로 연관된 수치 데이터를 생산하기 때문에 실제 산업 현장에서는 유용하겠지만, 과학 연구에 사용하는 목적으로는 적절하지 않다고 판단하였다.

3.1.2. EDA 및 전처리

Train set에서 진행한 EDA 단계에서 단백질 구조만으로 각 아미노산의 개수, 비율과 tm 변수 간의 상관관계를 확인했을 때 소수성(비극성) 아미노산이 차지하는 비율이 높으면 tm이 증가하고, 친수성(극성) 아미노산의 비율이 낮으면 tm이 감소하는 등 몇 가지 특이점을 찾아냈지만, Test set에서는 하나의 야생형 효소(wild type)를 기준으로 모든 데이터가 하나의 아미노산만이 바뀌고 있기 때문에 개수와 비율 변화는 예측에 거의 영향을 미치지 못할 것이 확실시되었다. 또한 EDA 과정에서 단일 아미노산 변형에 따라 크게 달라지는 정보를 찾을 수 없었기 때문에, 우리는 데이터셋의 구조를 바꾸고, 외부로부터 이러한 변형을 잘 반영할 수 있는 변수를 불러와 새로운 훈련 데이터를 생성할 계획을 세웠다.

3.1.3. 새로운 훈련 데이터 생성

Train set은 길이와 구조가 다른 여러 단백질과 그에 따른 tm값이 주어졌지만, Test set은 하나의 야생형 효소와 효소 구조에서 1개의 아미노산만이 변형된 단일 돌연변이들에 대한 tm값이 주어져 있다. 따라서 예측 정확도를 위해 Train set을 Test set과 비슷하게 조정하는 작업이 필요했다. 우선 원본 Train set에서 아미노산의 단일 변이체 데이터를 찾아 그룹화하였고, 그룹 내 데이터가 5개 이상인 그룹만을 추출해 총 78개의 유의미한 그룹을 생성하였다.

또한 Test set 내 모든 단백질의 pH와 데이터 출처가 동일하기 때문에 이와 비슷한 환경을 Train set에서 구축하기 위한 필터링을 진행했다. Train set의 각 그룹에서 가장 빈도수가 높은 pH를 찾아 그에 해당하는 단백질을 필터링했다. 데이터 출처는 필터링 기준으로 포함했을 때 추출된 데이터 수가 많이 감소했기 때문에 고려하지 않았다. 다음으로 단백질 구조 예측 프로그램인 AlphaFold의 데이터베이스에 각 그룹의 단백질 정보를 검색하여, 데이터베이스 내 단백질과 완전히 일치하는 단백질이 존재하는 그룹을 필터링하고 해당 단백질은 Test set과 같이 야생형 효소(wild type)로 지정하여 최종 Train set을 완성하였다.

최종 Train set은 73개의 그룹으로 구성되었으며 2556개의 단일 변이 구조가 담겨있다. 또한 캐글은 Test set의 야생형 효소에 대한 AlphaFold 예측 파일을 제공하였지만 Train set에 존재하는 효소에 대해서는 그러한 파일을 제공하지 않았기 때문에, 찾아낸 Train set의 73개의 야생형 효소에 대한 AlphaFold 예측 파일을 다운로드하여 학습에 활용할 계획을 세웠다. 그 이유는 AlphaFold가 제공하는 ‘B-factor’ 때문인데, 이 값의 의미는 3.1.4. 데이터 특성 선택에서 설명하도록 하겠다.

	PDB	WT	position	MUT	dTm	sequence	mutant_seq
0	GP01	L	89	A	2.28642	MLVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	MLVMTEYLLSAGICMAI
1	GP01	T	95	C	1.48642	MLVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	MLVMTEYLLSAGICMAI
2	GP01	T	95	C	0.28642	MLVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	MLVMTEYLLSAGICMAI
3	GP01	T	95	S	2.48642	MLVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	MLVMTEYLLSAGICMAI
4	GP01	T	95	S	3.88642	MLVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	MLVMTEYLLSAGICMAI

equence	mutant_seq	CIF
LVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	MLVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	AF-P00644-F1
LVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	MLVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	AF-P00644-F1
LVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	MLVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	AF-P00644-F1
LVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	MLVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	AF-P00644-F1
LVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	MLVMTEYLLSAGICMAIVSILLIGMAISNVSKGQYAKRFFFFATSC...	AF-P00644-F1

그림 2. new_train.csv

3.1.4. 데이터 특성 선택

데이터 특성으로는 크게 두 가지를 선택했다. 첫 번째는 아미노산 정보이고 두 번째는 ESM Transformer로 찾아낸 특성이다.

변이 전/후 아미노산 정보 관련 특성

훈련 데이터 내부에서는 변화하는 아미노산 각각에 대한 정보를 수집할 수 없어 외부 데이터를 이용해 아미노산의 물리적, 화학적 특성을 고려했다. 현재 존재하는 22가지 아미노산 중 단백질 구성에 이용되는 20가지 아미노산에 대한 특성을 고려했으며, 대표적으로 아미노산의 무게, 잔기, 등전점, 소수성, 극성, 편극률, 용제 접근 가능 표면적 등이 있다. 추가로 변이된 자리의 아미노산이 원래 야생형 효소에서 어떤 아미노산이었는지 나타내는 변수('AA1','AA2'), 변이된 위치의 아미노산 바로 앞과 뒤에 결합한 아미노산이 무엇인지에 대한 정보를 담고 있는 변수('AA3','AA4')도 생성하였다. 다만 어떤 정보가 단백질 안정성과 관련되어 중요한지에 대한 지식이 부족했기 때문에 이 단계에서 생성된 모든 변수를 전부 훈련 데이터로 이용하여 모델이 직접 변수의 중요도를 판단하도록 했다.

ESM protein transformer

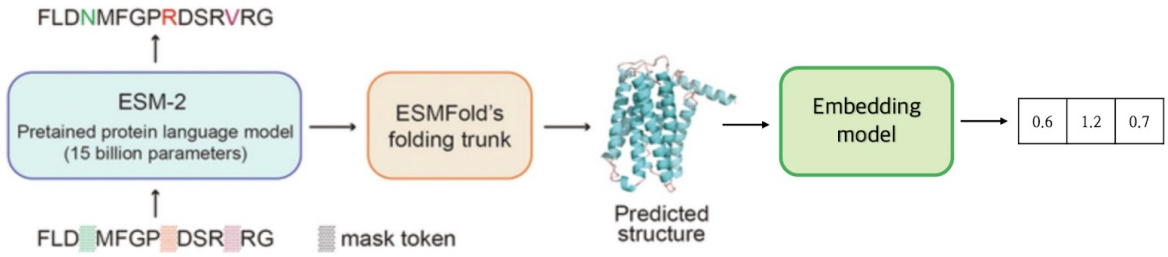

그림 3. ESM protein transformer

Train set은 단일 아미노산 돌연변이 외에도 각 그룹을 대표하는 야생형 효소(wild type)가 존재하기 때문에 기계학습을 위해서는 해당 효소 간의 변동이 수치 형태로 표현된 변수들이 필요했다. 이를 위해 우리는 메타 리서치(페이스북 AI 리서치)에서 공개한 단백질 언어 변환 모델인 ESM-2를 이용하여 단백질 구조 정보를 수치화된 변수로 생성하기로 했다. 우선 Train set 내 야생형 효소를 임베딩한 결과 전체를 'all_pdb_embed_local'에 저장하였다. 그리고 자연어 처리에서 문맥 정보를 얻기 위해 사용하는 Average word embedding 방식과 비슷하게 그룹별로 산출된 결괏값을 평균화하여 'all_pdb_embed_pool'에 저장함으로써 단백질의 변화 전반에 대한 정보를 나타낼 수 있도록 했다. 한편 ESM-2는 자연어 처리에 기반을 둔 AI 모델로서 단백질 내의 미세한 구조 변화도 잘 인식하기 때문에 우리는 단일 돌연변이에 대한 더 많은 정보 또한 ESM-2를 통해 수치화하기로 했다. 이를 위해 야생형 효소에서 변이되는 아미노산 위치에 해당하는 임베딩 정보를 'all_pdb_embed_tmp'에 저장했다. 산출된 데이터의 차원이 1280으로 너무 컸기 때문에 PCA를 진행하여 차원을 'all_pdb_embed_pool'은 32, 나머지는 16등으로 적절히 감소시켰다. 이와 같은 과정을 단일 아미노산 변이체 구조에도 적용하여 최종적으로 만들어진 변수는 pca_pool0 ~ 31, pca_local0~15, local 정보를 야생형(wildtype)과 돌연변이(mutant)로 구분한 pca_wt0 ~ 15, pca_mutant0 ~ 15로 총 80가지이다.

PDB 파일에 포함된 안정성 관련 특성

AlphaFold는 단백질의 3차원 구조를 예측하고 그 구조에서 파생된 단백질의 여러 물리적 특성이 담긴 PDB 파일을 생성한다. 이 특성들 중 우리는 B-factor에 주목했다. 원래 B-factor는

분자 내에서 특정 원자의 열운동성을 측정하는 단위로 우리 연구의 목표인 tm과 직접적인 상관관계가 있지만, AlphaFold는 구조를 단순히 예측할 뿐 측정하는 도구가 아니기 때문에 다른 방식을 통해 B-factor와 유사한 값을 계산하였으며 이를 pLDDT라 한다. pLDDT는 예측된 단백질 분자 내 각 아미노산의 3차원 좌표상 위치에 관한 신뢰도가 0~100 사이의 수치로 표현된 결과이다. 높은 pLDDT는 AlphaFold가 예측한 해당 아미노산의 위치, 해당 아미노산 앞/뒤로 연결된 결합 구조가 실제와 비슷하게 잘 추정됨을 의미한다. 현재 pLDDT가 단백질 분자의 안정성과 의미상으로 연관되어 있는지는 밝혀지지 않았지만, 우리는 안정된 단백질일수록 결합 구조 또한 안정되고 견고하기 때문에 pLDDT가 높아진다는 가설을 세웠고 이 값을 모델 학습을 위한 변수로 추가하기로 했다.

'tm' 변수를 정규화된 순위 변수로 변환

tm의 순위를 예측하는 우리의 목적에 맞게, 섭씨온도 단위인 기존 tm 변수에 그룹별로 순위를 매긴 새로운 변수를 목표변수로 지정했다. 높은 tm을 기록한 단백질은 높은 온도에서도 안정적이기 때문에 높은 순위가 부여되었다. 그리고 그룹별로 데이터의 수가 달라 순위의 범위가 그룹별로 달라지는 문제를 해결하기 위해 모든 순위를 그룹 내 데이터 수로 나누어 정규화하였다. 이 작업을 거쳐 모든 순위 변수는 0~1 사이에 위치하게 된다.

	WT	MUT	position	relative_position	b_factor	Molecular Weight_1	Molecular Weight_2	Molecular Weight_delta	Residue Weight_1	Residue Weight_2	Residue Weight_delta
0	L	A	89	0.385281	95.07	131.18	89.10	-42.08	113.16	71.08	-42.08
1	T	C	95	0.411255	98.41	119.12	121.16	2.04	101.11	103.15	2.04
2	T	C	95	0.411255	98.41	119.12	121.16	2.04	101.11	103.15	2.04
3	T	S	95	0.411255	98.41	119.12	105.09	-14.03	101.11	87.08	-14.03
4	T	S	95	0.411255	98.41	119.12	105.09	-14.03	101.11	87.08	-14.03

H_1	H_2	H_delta	VSC_1	VSC_2	VSC_delta	P1_1	P1_2	P1_delta	P2_1	P2_2	P2_delta	SASA_1	SASA_2
1.06	0.62	-0.44	93.5	27.5	-66.0	4.9	8.1	3.2	0.186	0.046	-0.140	1.931	1.181
-0.05	0.29	0.34	51.3	44.6	-6.7	8.6	5.5	-3.1	0.108	0.128	0.020	1.525	1.461
-0.05	0.29	0.34	51.3	44.6	-6.7	8.6	5.5	-3.1	0.108	0.128	0.020	1.525	1.461
-0.05	-0.18	-0.13	51.3	29.3	-22.0	8.6	9.2	0.6	0.108	0.062	-0.046	1.525	1.298
-0.05	-0.18	-0.13	51.3	29.3	-22.0	8.6	9.2	0.6	0.108	0.062	-0.046	1.525	1.298

SASA_delta	NCISC_1	NCISC_2	NCISC_delta	prev	b_factor_prev	post	b_factor_post	cos_angle	location3d
-0.750	0.051672	0.007187	-0.044485	K	88.59	H	95.97	-0.663945	137.210421
-0.064	0.003352	-0.036610	-0.039962	A	98.61	L	98.69	-0.828820	328.740432
-0.064	0.003352	-0.036610	-0.039962	A	98.61	L	98.69	-0.828820	328.740432
-0.227	0.003352	0.004627	0.001275	A	98.61	L	98.69	-0.828820	328.740432
-0.227	0.003352	0.004627	0.001275	A	98.61	L	98.69	-0.828820	328.740432

pca_pool_0	pca_wt_0	pca_mutant_0	pca_local_0	pca_pool_1	pca_wt_1	pca_mutant_1	pca_local_1	pca_pool_2
-0.014211	3.889750	4.110106	0.220357	-0.007354	0.857227	0.407647	-0.449580	0.015841
0.048098	2.957542	3.031531	0.073989	-0.001994	-0.031811	0.216848	0.248658	-0.027369
0.048098	2.957542	3.031531	0.073989	-0.001994	-0.031811	0.216848	0.248658	-0.027369
0.021657	2.957542	2.995182	0.037640	-0.005534	-0.031811	-0.107469	-0.075658	-0.007399
0.021657	2.957542	2.995182	0.037640	-0.005534	-0.031811	-0.107469	-0.075658	-0.007399

pca_pool_29	pca_pool_30	pca_pool_31	AA1	AA2	AA3	AA4	ddG	dTm	pdb	source	target
-0.004637	-0.007803	-0.001802	9	0	8	6	NaN	2.28642	GP01	kaggle.csv	0.578189
-0.004492	-0.007832	-0.015564	16	1	0	9	NaN	1.48642	GP01	kaggle.csv	0.512346
-0.004492	-0.007832	-0.015564	16	1	0	9	NaN	0.28642	GP01	kaggle.csv	0.425926
0.006524	0.011033	0.015508	16	15	0	9	NaN	2.48642	GP01	kaggle.csv	0.592593
0.006524	0.011033	0.015508	16	15	0	9	NaN	3.88642	GP01	kaggle.csv	0.769547

그림 4. Feature Engineering 이후 Train set

3.1.5. 모델링

Test set이 하나의 돌연변이 그룹으로 이루어져 있고 훈련 데이터 또한 이에 맞추어 그룹화되어 있기 때문에, 그룹 변수를 Cross-validation에 반영하는 GroupKFold를 적용하였다(k=10). 다만 Train set에 포함된 73개의 그룹으로 GroupKFold를 적용하면 Validation set에 포함되는 그룹이 너무 많아지기 때문에, 전체 데이터 크기를 크게 감소시키지 않는 선에서 데이터가 적은 그룹을 학습에서 추가적으로 제외하였다. 이에 따라 총 31개의 그룹만이 남게 되었다. 학습에 사용된 모델은 2장에서 소개한 XGBoost, LightGBM, Ridge, Random Forest를 사용했으며, Grid Search로 hyper parameter tuning을 진행하였다. 변수 생성 과정을 거친 총 127개의 변수가 학습에 이용되었다.

```
We have 127 features for our model:
['position', 'relative_position', 'b_factor', 'Molecular Weight_1', 'Molecular Weight_2',
'Molecular Weight_delta', 'Residue Weight_1', 'Residue Weight_2', 'Residue Weight_delta',
'pKa1_1', 'pKa1_2', 'pKa1_delta', 'pKb2_1', 'pKb2_2', 'pKb2_delta', 'pKx3_1', 'pKx3_2', 'pK
x3_delta', 'pl4_1', 'pl4_2', 'pl4_delta', 'H_1', 'H_2', 'H_delta', 'VSC_1', 'VSC_2', 'VSC_d
elta', 'P1_1', 'P1_2', 'P1_delta', 'P2_1', 'P2_2', 'P2_delta', 'SASA_1', 'SASA_2', 'SASA_de
lta', 'NCISC_1', 'NCISC_2', 'NCISC_delta', 'b_factor_prev', 'b_factor_post', 'cos_angle',
'location3d', 'pca_pool_0', 'pca_wt_0', 'pca_mutant_0', 'pca_local_0', 'pca_pool_1', 'pca_w
t_1', 'pca_mutant_1', 'pca_local_1', 'pca_pool_2', 'pca_wt_2', 'pca_mutant_2', 'pca_local_
2', 'pca_pool_3', 'pca_wt_3', 'pca_mutant_3', 'pca_local_3', 'pca_pool_4', 'pca_wt_4', 'pca
_mutant_4', 'pca_local_4', 'pca_pool_5', 'pca_wt_5', 'pca_mutant_5', 'pca_local_5', 'pca_po
ol_6', 'pca_wt_6', 'pca_mutant_6', 'pca_local_6', 'pca_pool_7', 'pca_wt_7', 'pca_mutant_7',
'pca_local_7', 'pca_pool_8', 'pca_wt_8', 'pca_mutant_8', 'pca_local_8', 'pca_pool_9', 'pca_
wt_9', 'pca_mutant_9', 'pca_local_9', 'pca_pool_10', 'pca_wt_10', 'pca_mutant_10', 'pca_loc
al_10', 'pca_pool_11', 'pca_wt_11', 'pca_mutant_11', 'pca_local_11', 'pca_pool_12', 'pca_wt
_12', 'pca_mutant_12', 'pca_local_12', 'pca_pool_13', 'pca_wt_13', 'pca_mutant_13', 'pca_lo
cal_13', 'pca_pool_14', 'pca_wt_14', 'pca_mutant_14', 'pca_local_14', 'pca_pool_15', 'pca_w
t_15', 'pca_mutant_15', 'pca_local_15', 'pca_pool_16', 'pca_pool_17', 'pca_pool_18', 'pca_p
ool_19', 'pca_pool_20', 'pca_pool_21', 'pca_pool_22', 'pca_pool_23', 'pca_pool_24', 'pca_po
ol_25', 'pca_pool_26', 'pca_pool_27', 'pca_pool_28', 'pca_pool_29', 'pca_pool_30', 'pca_poo
l_31', 'AA1', 'AA2', 'AA3', 'AA4']
```

그림 5. Result of Feature selection

3.2. 실험 결과

3.2.1 Feature Importance (What is the most important feature?)

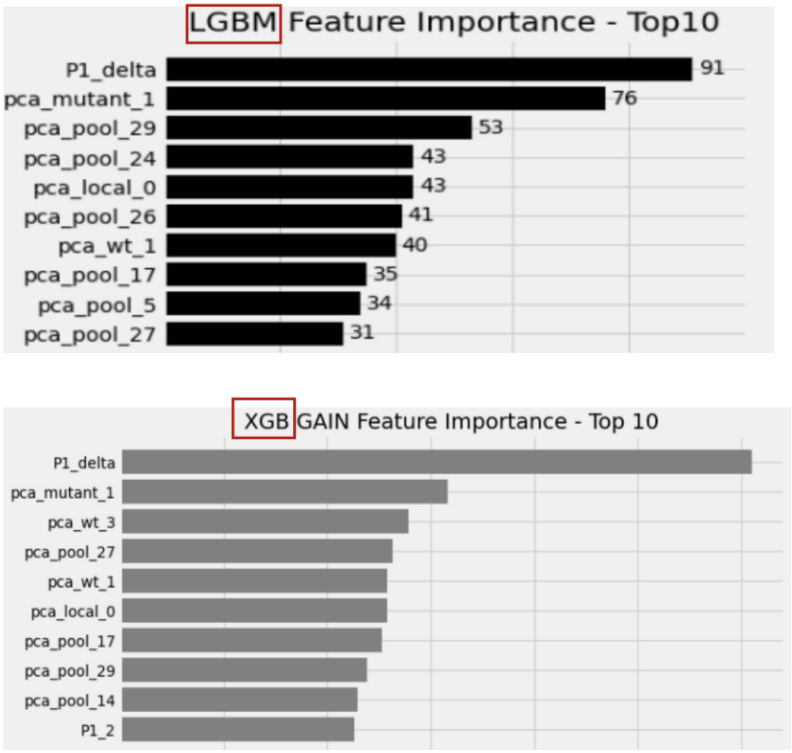

그림 6. Feature Importance

LGBM, XGB 모델에서 상위 10개의 변수 중요도를 추출하였고 두 가지 사실을 확인할 수 있었다. 첫 번째, 공통으로 P1_delta가 모델 생성에 가장 중요한 변수로 나타났다. 이는 아미노산의 극성에 관련된 변수로 Wildtype과 Mutation 아미노산 사이의 극성 차이를 나타낸 것이다. 이를 통해 단일 아미노산 변형에 따른 극성 변화가 안정성에 큰 영향을 미침을 알 수 있다. 두 번째, ESM-2에 의해 생성된 변수들의 중요도가 높다는 것이다. 이것은 단백질 구조를 수치화시킨 뒤 PCA를 적용한 변수들로, 따라서 전체적인 단백질 구조 상에서의 미지의 변화가 모델링에 미치는 영향이 큰 것을 확인할 수 있었다. 실제로 단백질은 일부 아미노산이 변형될 때 3차원적인 단백질 접힘 형태가 변화하게 되고 이러한 3차원적인 구조 변화가 단백질 안정성의 변화와 관련되어 있다고 알려져 있다.

3.2.2 Best Model

본 대회는 캐글에서 진행되었기 때문에 리더보드 점수인 스피어만 순위상관계수를 평가 기준으로 설정했다. 모델링 결과 XGB:0.273, LGBM:0.291, RF:0.264, Ridge:0.255 점이 나왔다. 따라서 가장 좋은 성능을 가진 모델은 LGBM이다. 전체 대회 참가자 중 최고점은 0.859점으로 우리의 점수는 높은 편이 아니지만, 현재 대회에서 0.5 이상의 상관계수를 기록한 결과는 대부분 Rosetta, DeepDGG 등 단백질 변형에 의한 안정성 변화를 직접적으로 예측하는 외부 모델의 사용과 Train set 외의 추가적인 단백질 데이터의 사용을 극대화한 경우이다. 우리는 오직 Train set만을 훈련 데이터로 사용했고 캐글에서 제공한 AlphaFold의 pLDDT 예측값 외에는 단백질 안정성과 연관된 특성을 훈련에 적용하지 않았기 때문에 우리가 얻은 최종 점수 0.291은 상당히 괜찮은 결과라 여겨진다. 실제로 원본 Train set에서 LightGBM의 단순 적용을 실시한 참가자의 최종 상관계수는 0.052에 불과했으며 우리의 특성 연구가 성공적이었음을 간접적으로 방증하고 있다.

4. 결론

우리는 다양한 알고리즘을 사용하여 모델링을 진행하였다. 그 결과 일반적으로 성능이 좋다고 알려진 순서대로 리더보드 점수를 획득했다. 이는 우리의 데이터가 완전한 정형 모델이었기 때문이라고 생각된다. 또한 특성의 중요도를 살펴볼 때 아미노산 변형에 따른 극성 변화와 단백질 구조 상의 변화가 단백질 안정성에 영향을 미치는 것을 알 수 있었다. 모델링을 진행하며 아쉬웠던 점은 Feature Extraction에 집중했기에 앙상블 기법을 적용해서 성능을 끌어올리지 못한 점이었다. LinearRegression, XGBoost, CatBoost, Ridge로 앙상블을 진행했을 때 단독으로 사용한 LightGBM보다 성능이 좋지 않았기에 추후에 더 많은 실험을 통해 성능을 끌어올릴 수 있을 것이라 기대한다. 또한 PCA 적용 전후로 모델링을 진행해서 결과를 비교한다면 PCA로 중요한 정보에 대한 손실을 예방할 수 있기 때문에 후속 연구로 보완할 점이다.