

Tabular Playground Series

– Dec 2021

권도근

-담당: 프로젝트개요,모델,결론

김태용

-담당: 환경설정 및워크플로우,
데이터 전처리

목차 A table of contents.

01 프로젝트의 개요 및 목적

02 환경설정 및 워크플로우

03 데이터 수집 및 전처리

04 모델

05 프로젝트 결론



Part 1.

프로젝트의 개요 및 목적

Part 1, 캐글대회 소개



Playground Prediction Competition

Tabular Playground Series - Dec 2021

Practice your ML skills on this approachable dataset!

Kaggle · 1,097 teams · 3 days to go

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Prediction](#)

Overview	
Description	Kaggle competitions are incredibly fun and rewarding, but they can also be intimidating for people who are relatively new in their data science journey. In the past, we've launched many Playground competitions that are more approachable than our Featured competitions and thus, more beginner-friendly.
Evaluation	
Timeline	In order to have a more consistent offering of these competitions for our community, we're trying a new experiment in 2021. We'll be launching month-long tabular Playground competitions on the 1st of every month and continue the experiment as long as there's sufficient interest and participation.
Prizes	The goal of these competitions is to provide a fun, and approachable for anyone, tabular dataset. These competitions will be great for people looking for

Part 1, 프로젝트 개요

001 >> 데이터 불러오기

002 >> EDA & 데이터 전처리

003 >> LightGBM , XGBOOST , Tensor Flow
모델 구축

Part 1, 데이터 셋 설명

콜로라도 북부
주르벨트 국유림 단위
넓이: 30m*30m

종속변수: Cover_Type(구성
목)

- 가문비/전나무
- 로지폴 소나무
- 폰데로사 소나무
- 미루나무/버드나무
- 사시나무
- 미송
- Krummholz



Data

Elevation: 고도(m)
Aspect: 경사 방위면(degree)
Slope: 경사도(degree)

Horizontal_Distance_To_Hydrology: 지표수까지의 수평 거리(m)

Vertical_Distance_To_Hydrology: 지표수까지의 수직 거리(m)

Horizontal_Distance_To_Roadways: 도로까지의 수평 거리(m)

Hillshade_9am: 하지 오전 9시 음영도(0 to 255 index)

Hillshade_Noon: 하지 정오 음영도(0 to 255 index)

Hillshade_3pm: 하지 오후 3시 음영도(0 to 255 index)

Horizontal_Distance_To_Fire_Points: 산불 점화 지점까지의 수평 거리(m)

Wilderness_Area: 야생 지역(1~4, 0 또는 1 data)

1. Rawah
 2. Neota
 3. Comanche Peak
 4. Cache la Poudre
- Soil_Type: 토양 유형(1~40, 0 또는 1 data)

Part 2.

환경설정 및 워크플로우

Part 1, 환경 설정 및 워크플로우

프로젝트 주요 라이브러리 버전

라이브러리	버전	라이브러리	버전
Python	3.7.12	Xgboost	1.5.1
Numpy	1.19.5	Catboost	1.0.3
Pandas	1.3.4	Lightgbm	3.3.1
Plotly	5.4.0	Tenserflow	2.6.2
Sklearn	0.0		
Scipy	1.7.2		

Part 1, 환경 설정 및 워크플로우

- 딥러닝은 필수적으로 GPU를 사용해야 함
- GPU 무료 사용 – Google Colab, Kaggle Notebook
- 로컬 구축 – 리눅스 환경 또는 M1 Mac (본 연구 진행)
- 딥러닝 프레임워크
 - Tensorflow: <https://www.tensorflow.org/>
 - Pytorch: <https://pytorch.org/>

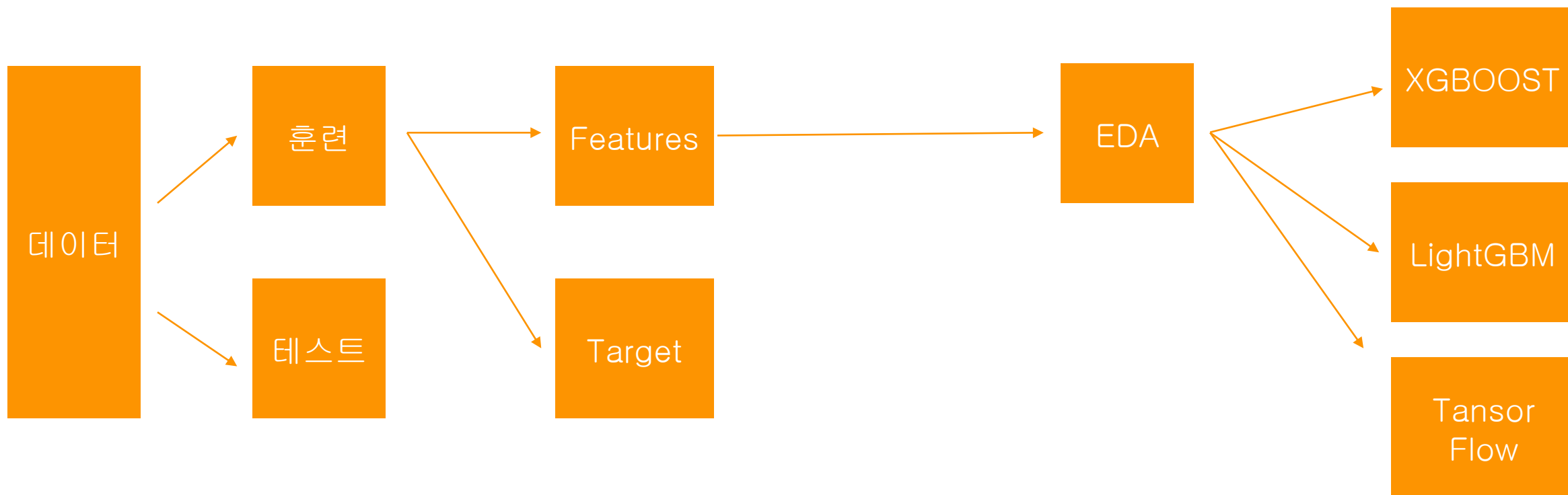
순서도

Train과 Test데이터로 분리

Target과 Features 데이터 분리

EDA

모델링



Part 3

데이터 수집 및 전처리

Part 1, 데이터 불러오기

Train Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4000000 entries, 0 to 3999999
Data columns (total 56 columns):
#   Column                                          Dtype
---  -
0   Id                                             int64
1   Elevation                                     int64
2   Aspect                                       int64
3   Slope                                         int64
4   Horizontal_Distance_To_Hydrology             int64
5   Vertical_Distance_To_Hydrology               int64
6   Horizontal_Distance_To_Roadways              int64
7   Hillshade_9am                               int64
8   Hillshade_Noon                              int64
9   Hillshade_3pm                               int64
10  Horizontal_Distance_To_Fire_Points            int64
...
47  Soil_Type33                                  int64
48  Soil_Type34                                  int64
49  Soil_Type35                                  int64
50  Soil_Type36                                  int64
51  Soil_Type37                                  int64
52  Soil_Type38                                  int64
53  Soil_Type39                                  int64
54  Soil_Type40                                  int64
55  Cover_Type                                    int64
dtypes: int64(56)
memory usage: 1.7 GB
None
```

Test Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000000 entries, 0 to 999999
Data columns (total 55 columns):
#   Column                                          Non-Null Count  Dtype
---  -
0   Id                                             1000000 non-null int64
1   Elevation                                     1000000 non-null int64
2   Aspect                                       1000000 non-null int64
3   Slope                                         1000000 non-null int64
4   Horizontal_Distance_To_Hydrology             1000000 non-null int64
5   Vertical_Distance_To_Hydrology               1000000 non-null int64
6   Horizontal_Distance_To_Roadways              1000000 non-null int64
7   Hillshade_9am                               1000000 non-null int64
8   Hillshade_Noon                              1000000 non-null int64
9   Hillshade_3pm                               1000000 non-null int64
10  Horizontal_Distance_To_Fire_Points            1000000 non-null int64
...
50  Soil_Type36                                  1000000 non-null int64
51  Soil_Type37                                  1000000 non-null int64
52  Soil_Type38                                  1000000 non-null int64
53  Soil_Type39                                  1000000 non-null int64
54  Soil_Type40                                  1000000 non-null int64
dtypes: int64(55)
memory usage: 419.6 MB
None
```

Part 1, 데이터 수집 및 전처리

Total Data (Train + Test)

	count	mean	std	min	25%	50%	75%	max
Soil_Type27	5000000.000000	0.011771	0.107856	0.000000	0.000000	0.000000	0.000000	1.000000
Soil_Type21	5000000.000000	0.011586	0.107011	0.000000	0.000000	0.000000	0.000000	1.000000
Soil_Type9	5000000.000000	0.010968	0.104150	0.000000	0.000000	0.000000	0.000000	1.000000
Soil_Type28	5000000.000000	0.010767	0.103203	0.000000	0.000000	0.000000	0.000000	1.000000
Soil_Type36	5000000.000000	0.010709	0.102930	0.000000	0.000000	0.000000	0.000000	1.000000
Soil_Type3	5000000.000000	0.009225	0.095602	0.000000	0.000000	0.000000	0.000000	1.000000
Soil_Type25	5000000.000000	0.003276	0.057146	0.000000	0.000000	0.000000	0.000000	1.000000
Soil_Type8	5000000.000000	0.002921	0.053971	0.000000	0.000000	0.000000	0.000000	1.000000
Soil_Type15	5000000.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Soil_Type7	5000000.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Soil_Type(토양 유형): 해당하는 토양 유형일 경우 1의 값이 입력된 columns

- 15, 7: 측정 data가 존재하지 않아 해당 column 제거
- 3, 8, 25: 측정 data가 전체의 1% 미만인 column 제거

Part 1, 데이터 수집 및 전처리

Train Data

	Cover_Type	count
0	1	1468136
1	2	2262087
2	3	195712
3	4	377
4	5	1
5	6	11426
6	7	62261



	Cover_Type	count
0	1	1468136
1	2	2262087
2	3	195712
3	6	11426
4	7	62261

Cover_Type: 30 x 30 (m²) 을 구성하는 주요 나무

- **Type 4:** 미루나무/버드나무
7개 Type 중 Train 전체 data(400000개)의 0.1% 미만을 차지한다.
- **Type 5:** 사시나무
단 1개의 Raw만 포함된 Type으로 학습하기 적절하지 않다고 판단.

Part 4.

모델링

Part 1, 모델 알고리즘 소개

XGBOOST

- 트리 기반의 앙상블 학습
- 캐글 경연에서 상위 데이터에서 xgboost사용
- GBmd에 기반하지만 느린 수행시간 및 과적합 규제 부재 문제 보완

LightGBM

- LightGBM은 리프 중심 트리 분할 방식
- LightGBM의 리프 중심 트리 분할은 트리의 균형을 맞추지 않고 최대 손실 값을 가지는 리프 노트를 지속적으로 분할
- 균형 트리의 분할 방식보다 예측 오류 손실을 최소화 가능

Tensor Flow

- 머신러닝을 위한 신경망을 쉽게 빌드할 수 있도록 설계
- 텐서플로우(TensorFlow)는 기계 학습과 딥러닝을 위해 구글에서 만든 오픈소스 라이브러리

Part 1, XGBOOST vs LightGBM

———— XGBOOST와 LightGBM은 부스팅 기법 앙상블 알고리즘과 관련

앙상블이란?

여러 개의 결정 트리(Decision Tree)를 결합하는 것으로, 하나의 결정 트리보다 알고리즘 성능을 더 높임

앙상블 학습을 통해 약한 분류기(Weak Classifier) 여러 개를 결합해서 강한 분류기(Strong Classifier)를 만들 수 있음

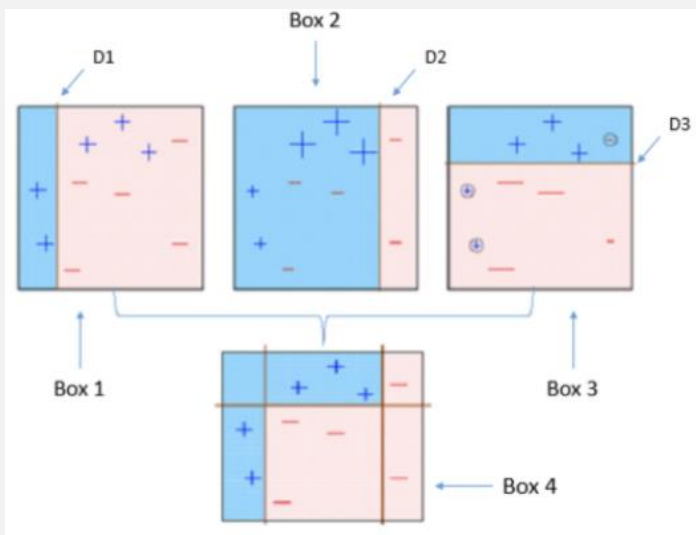
머신러닝 앙상블은 크게 배깅(Bagging) 부스팅(Boosting)으로 구분

Part 1, XGBOOST vs LightGBM

XGBOOST와 LightGBM은 부스팅 기법 앙상블 알고리즘과 관련

부스팅이란?

부스팅(Boosting)은 이전 모델의 예측 결과에 따라 가중치를 활용해서 강분류기를 만드는 방법



예시

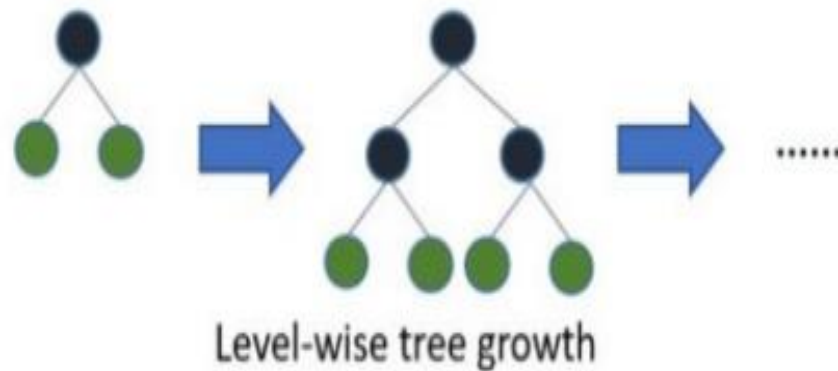
그림은 +와 -로 구성된 데이터셋을 분류하는 문제

잘못 분류된 데이터는 가중치를 높여주고, 잘 분류된 데이터는 가중치를 낮추어 주며 데이터 크기를 조절하는 과정을 통해 다음 모델에서 해당 데이터에 더 집중해 분류되고 있다

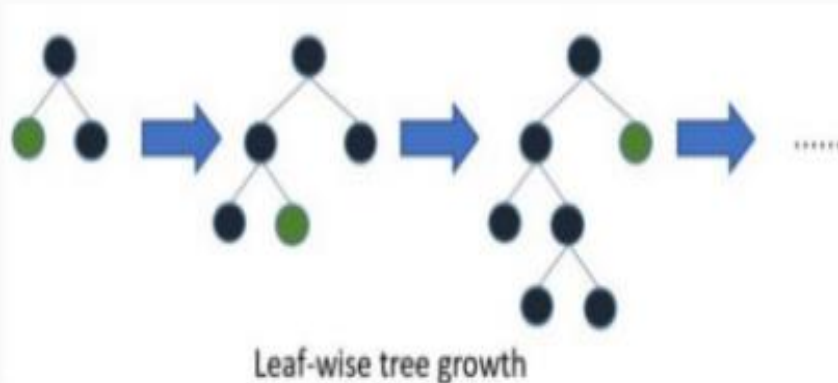
D1, D2, D3의 classifier를 합쳐 최종 classifier를 구할 수 있다

Part 1, XGBOOST vs LightGBM

XGBoost:



LightGBM:



XGBOOST

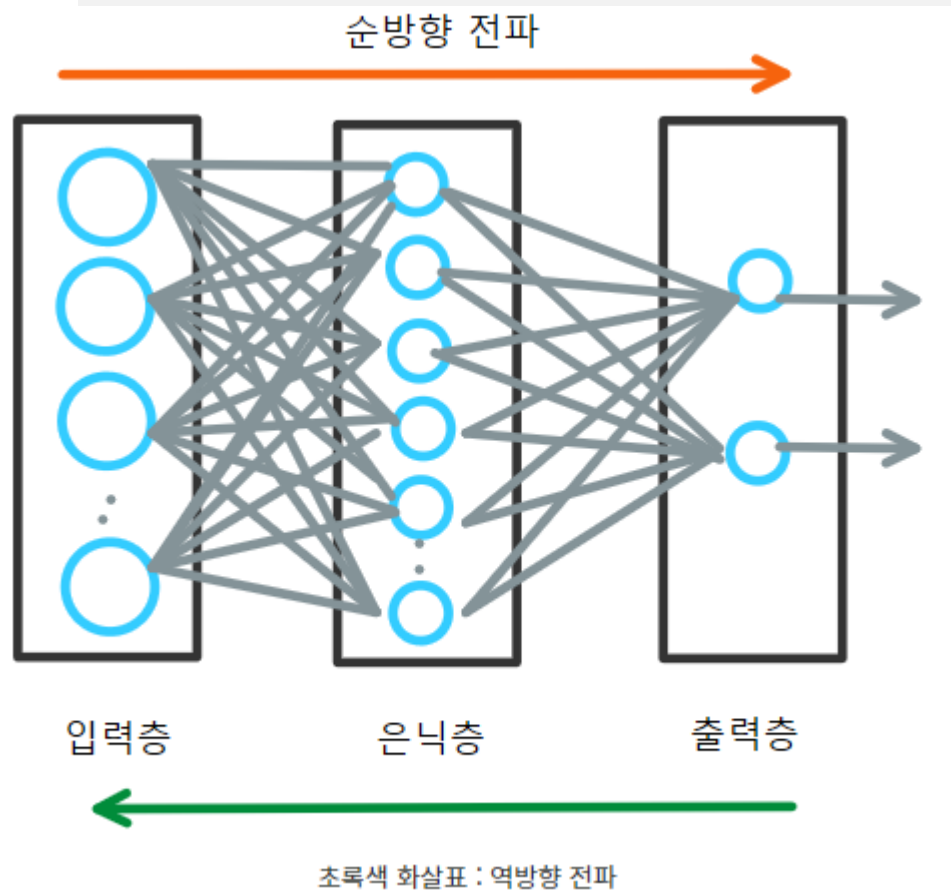
- 이전 GBM보다 성능은 좋아졌지만 여전히 학습시간이 느림

LightGBM

- LightGBM은 대용량 데이터 처리가 가능하고 메모리를 적게 사용하며 빠름
- LGBM은 리프 노드를 중심으로 트리 분할
- LGBM은 균형은 상관없이 loss를 가장 줄일 수 있는 쪽으로 리프 노드를 지속적으로 분할

Part 1, Tensor Flow

MLP DeepLearning



- 여러층의 퍼셉트론으로 적어도 1개 이상의 은닉층(hidden layer) 보유
- 일반적으로 지도학습
- 역전파 알고리즘(Backpropagation)으로 학습 - 다층 퍼셉트론 문제 해결위한 알고리즘
- 경사하강법으로 에러를 최소화
- 복잡한 데이터의 분류가 가능

Part 1, Tensor Flow

Swish 함수

활성화 함수(Activation Function)는 입력을 받아 활성, 비활성을 결정하는데 사용되는 함수

기존 활성화 함수

Activation Functions

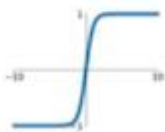
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



tanh

$$\tanh(x)$$



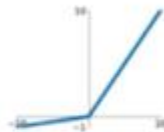
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

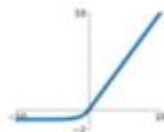


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Swish 함수

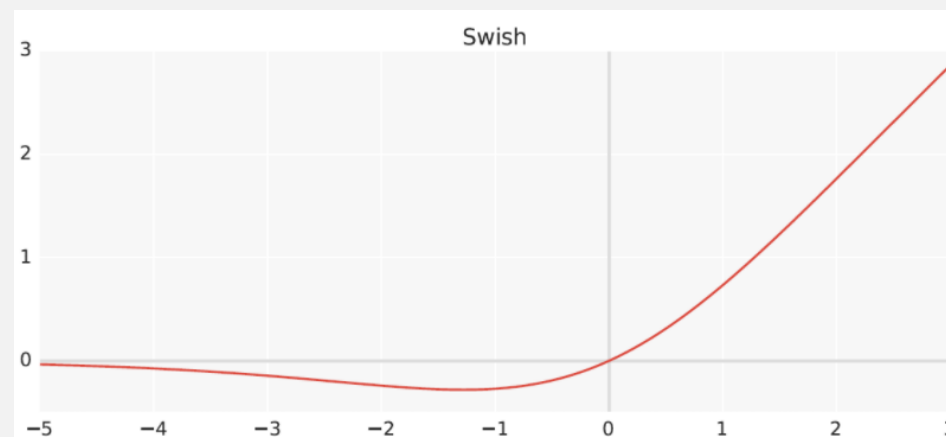


Figure 1: The Swish activation function.

- Swish 는 매우 깊은 신경망에서 ReLU 보다 높은 정확도를 달성
- 또한 모든 배치 크기에 대해 Swish 는 ReLU 를 능가
- 모든 $x < 0$ 에 대해 함수를 감소시키거나 증가시키지 않는다.
- Mish 와 마찬가지로 bounded below, unbounded above 특징을 가진다.

Part 1, 모델 평가

실험	Model	실험내용	근거	점수
실험 1	lightGBM	default		0.94071
실험 2	lightGBM	FEATURES['mean', 'std', 'min', 'max'] 사용하지 않음	단순히 모든 컬럼에서 평균, 중간, 최소, 최대값을 가져오는 columns 이기 때문에 ML 상에서 중요하지 않음	0.94345
실험 3	lightGBM	total DF에서 적은 점유율(1% 이하)을 보이는 Soild_Type 제거	임의로 뽑아낸 값이기 때문에 train과 test를 모두 고려	0.94545
실험 4	lightGBM	Cover_Type == 4 를 제거	표본이 적어 상관관계 영향이 적을거라 판단	0.95069
실험 5	CatBoost	실험4와 동일		0.95362
실험 6	XgBoost	실험4와 동일		0.95246
실험 7	Tensorflow	실험4와 동일		0.95519
실험 8	lightGBM	Cover_Type == 6 제거	1%미만의 수치	0.95056
실험 9	lightGBM	1.5% 미만인 Soil_Type 제거	상관관계의 기준치를 올림	0.94668
실험 10	lightGBM	0.5% 미만인 Soil_Type 제거	상관관계의 기준치를 내림	0.95035

Part 5.

프로젝트 결론

Part 1, 경연


Playground Prediction Competition

Tabular Playground Series - Dec 2021

Practice your ML skills on this approachable dataset!

Kaggle · 1,133 teams · 2 days to go

Overview Data Code Discussion Leaderboard Rules Team My Submissions **Submit Predictions** ...

369	hangack		0.95519	13	1h
-----	---------	---	---------	----	----

Your Best Entry ↑

Your submission scored 0.95484, which is not an improvement of your best score. Keep trying!

1133개의 팀 중 0.95519점수 369등