



# 모의경진대회 OT 및 1주차 Default Prediction 과제 특강

(주)마인즈앤컴퍼니 | 김태훈 매니저

# Index

1. 모의경진대회 개괄
2. AI CONNECT 플랫폼 소개
3. 과제 - Default Prediction
4. 평가지표 - Macro F1 Score
5. 모델 - XGBoost & Ensemble

# ML / DS 기본 교육

## 교육목표

- 머신러닝/딥러닝에 대한 기초 이론 및 용어를 이해할 수 있다.
- 이미지/자연어/수치해석 분야 데이터에 대한 특징을 이해하고 처리할 수 있다.
- 분야별 알고리즘을 이해하고 관련 라이브러리 사용법을 이해할 수 있다.

## 교육 특징

- 교수님과 전문가를 통한 기본 이론 및 실습 교육 병행
- Python 기반의 라이브러리, 딥러닝 프레임 워크 실습 운영
- Google Colab 실습 환경에서 진행

## 진행 방식

- (오전 - 3시간) 주재걸 교수님이 주최하시는 이론 기본 교육 진행
- (오후 - 3시간) AI 멘토가 진행하는 이론/실습 과정 진행
- (오후 - 2시간) 실습에 따른 숙제 진행 및 풀이

교육명		세부내용
ML/DL 기초 교육	머신러닝 기초	<ul style="list-style-type: none"> <li>• 머신러닝 기초 알고리즘에 대하여 원리 위주의 이론 교육 실시</li> <li>• Linear Regression, Logistic Regression 등</li> <li>• 간단한 실습 진행</li> </ul>
	머신러닝 심화	<ul style="list-style-type: none"> <li>• 신경망 구조 및 딥러닝 알고리즘의 원리와 활용에 대한 이론 교육 실시</li> <li>• Neural Network, Backpropagation, Deep Neural Network 등</li> <li>• pytorch 사용법 안내 및 간단한 실습</li> </ul>
	수치해석 분야 인공지능 모델	<ul style="list-style-type: none"> <li>• 예측, 분류, 시계열 분석 등 수치해석 분야에 다양한 알고리즘에 대한 이론 교육 실시</li> <li>• SVM, LGBM, LSTM 등</li> <li>• 수치 해석 문제에 대한 접근 방법 및 앙상블 방법론 교육</li> </ul>
	이미지 분야 인공지능 모델	<ul style="list-style-type: none"> <li>• 이미지 분류, 객체인식, 영역탐지 등 이미지 분야에 다양한 알고리즘에 대한 이론 교육 실시</li> <li>• VGG, YOLO, Fine-tuning 기법 등</li> <li>• 이미지 처리 방법 및 다양한 평가지표에 대한 안내</li> </ul>
	자연어 분야 인공지능 모델	<ul style="list-style-type: none"> <li>• 문서 및 문장 분류, 감성분류, 기계독해, 문서 요약 등 다양한 알고리즘에 대한 이론 교육 실시</li> <li>• W2V, RNN, LSTM, Seq2Seq 등</li> <li>• 자연어 전처리 방법 및 다양한 평가지표에 대한 안내</li> </ul>

# 이론 심화 교육

## 교육목표

- 모의 캐글 대회 진행을 위한, AI 모델 심화 이론을 이해할 수 있다.
- 캐글 예제를 통하여 문제별 적절한 라이브러리 및 프레임워크 활용 할 수 있다.
- 모델 고도화를 위한 파라미터 튜닝 및 최적화를 수행 할 수 있다.

## 교육 특징

- ML/DL 심화 내용에 대해서 이론 및 실습 교육 진행
- Kaggle 예제 중심으로, 이론 및 실습 교육 진행
- Google Colab을 활용한 실습 환경

## 진행 방식

- (오전 - 3시간) 인공지능 심화 이론 교육
- (오후 - 3시간) Kaggle 사례 중심의 분야별 모델 학습 및 실습 진행
- (오후 - 2시간) 실습에 따른 숙제 진행 및 풀이

교육명		세부내용
캐글 기초 교육	캐글을 위한 Python 및 머신러닝 기초	<ul style="list-style-type: none"> <li>• 캐글을 진행하기 위한 실전Python 기초 과정 수행</li> <li>• 머신러닝 과제 수행을 위한 Pandas, Numpy 등 다양한 라이브러리 기능 교육</li> <li>• 데이터 시각화에 대한 기초를 습득하며, 캐글의 다양한 사례를 기반으로 교육</li> </ul>
캐글 심화 교육	캐글 진행을 위한 딥러닝 심화 교육 - 이미지/영상 -	<ul style="list-style-type: none"> <li>• 이미지 데이터 기반 Classification, Detection, Segmentation 등 다양한 테스크 수행 방법론 교육 및 실습</li> <li>• VGG, ResNet, YOLO, U-Net 등</li> </ul>
	캐글 진행을 위한 딥러닝 심화 교육 - 자연어 -	<ul style="list-style-type: none"> <li>• 텍스트 데이터 기반 Classification, QA 등 다양한 테스크 수행을 위한 방법론 교육 및 실습</li> <li>• W2V, RNN, Seq2seq, BERT 등</li> </ul>
	캐글 진행을 위한 딥러닝 심화 교육 - 수치해석 -	<ul style="list-style-type: none"> <li>• Numeric 데이터 기반 Classification, Time-series analysis 등 다양한 테스크 방법론 교육 및 실습</li> <li>• Boosting, Random Forest, SVM 등</li> </ul>

# 모의 경진대회 과정

## 교육목표

- 실전 캐글 대회 참여를 대비하여 모의 캐글 문제를 풀 수 있다.
- 이미지/자연어/수치해석 **Task** 별로, 적절한 모델을 사용하여 추론 및 결과 제출 가능하다.
- 경진대회 진행을 위한 **실험 설계 및 운영법**을 터득한다.

## 교육 특징

- 실전과 유사한 모의경진대회 통한 **실전 캐글에 대한 친숙도 증대**
- 모의 경진대회 참여를 통하여 **task별 AI 문제 해결 능력 향상**
- 참가 팀별 별도 서버 제공하여, 원활한 경진대회 참여를 위한 자원 제공

## 진행 방식

- **OT**를 통한 과제 설명 및 베이스라인 교육
- **개인전 1회 실시** / 개인전 성적에 따른 팀 매칭 및 **팀전 3회 실시**
- 멘토링 세션 통한 질의응답 및 문제 해결 지원

교육명		세부내용
모의 Kaggle 교육	개인전 (1회)	<ul style="list-style-type: none"> <li>• 금융 &amp; 정형 데이터 모의 Kaggle 경진대회</li> <li>• AI CONNECT 사용법, 과제, Baseline에 대한 기본 교육 실시</li> <li>• 실시간 리더보드 운영 및 개인 성적 산출</li> </ul>
	팀전 (3회)	<ul style="list-style-type: none"> <li>• 개인전 성적 통한 팀 매칭 및 팀별 서버 제공</li> <li>• 과제, Baseline에 대한 기본 교육 실시</li> <li>• 바이오 &amp; 이미지 데이터 모의 Kaggle 경진대회</li> <li>• 유통 &amp; 정형 데이터 모의 Kaggle 경진대회</li> <li>• 게임 &amp; 자연어 데이터 모의 Kaggle 경진대회</li> <li>• 실시간 리더보드 운영 및 팀별 성적 산출</li> </ul>

# 실전 경진대회 과정

## 교육목표

- 실전 경진대회 참여하여, 문제에 맞는 AI 모델 학습하여 결과물을 제출할 수 있다.
- 실전 대회에 맞는 실험을 설계 및 운영하여, 최적의 결과물을 제출할 수 있다.

## 교육 특징

- 기업 및 기관에서 출제한 실전 AI 경진대회에 참여하여 기술 역량을 강화
- 실전 대회 참여를 통한 AI 역량 증명 및 포트폴리오 작성
- 참가 팀별 서버 및 멘토링 지원

## 진행 방식

- 교육 OT를 통하여 문제 설명 및 베이스라인 교육 실시
- 별도 멘토링 세션 진행하여, 문의사항에 대한 지원 실시
- 팀별 대회 진행 및 정성평가를 위한 프레젠테이션 준비 작업 진행

교육명	세부내용
실전 경진대회 진행	<ul style="list-style-type: none"> <li>• 모의 경진대회 교육 이후, 실전 경진대회 준비</li> <li>• 캐글 및 Dacon 등 실전 경진대회 참여하여 성과 달성</li> </ul>

kaggle™

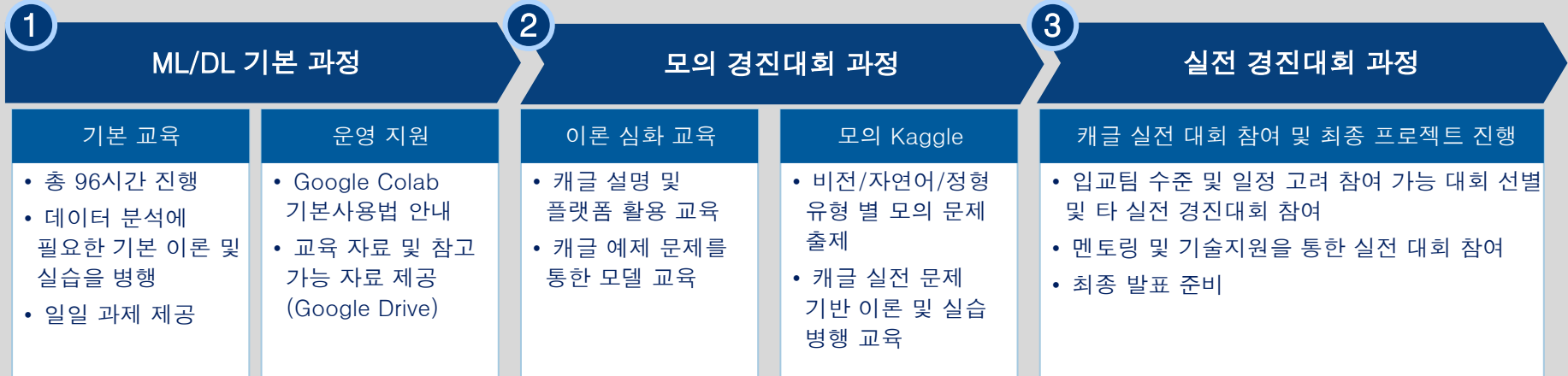
ai connect.

DACON

DATA TO VALUE

AI Factory

# 이론 교육부터 실전 대회 참여까지



## 3 경진대회 인프라 및 기술지원

### GPU 학습 서버

- 팀별 GPU서버 지원
- 실전 경진대회 운영을 위한 한시적 지원

### 기술 지원

- GPU 서버 통합 운영관리 및 유지보수 지원
- OS, Framework, Python 등 기본 프로그램 설치 제공

## 4 교육 및 멘토링 운영지원

### 교육

- 캐글 플랫폼 사용 실습 기본 교육
- 경진대회 문제 설명 및 baseline 코드 교육

### 멘토링

- KAIST AI 대학원 주재걸 교수 외 멘토 투입
- 공통 교육 및 팀별 멘토링 수행

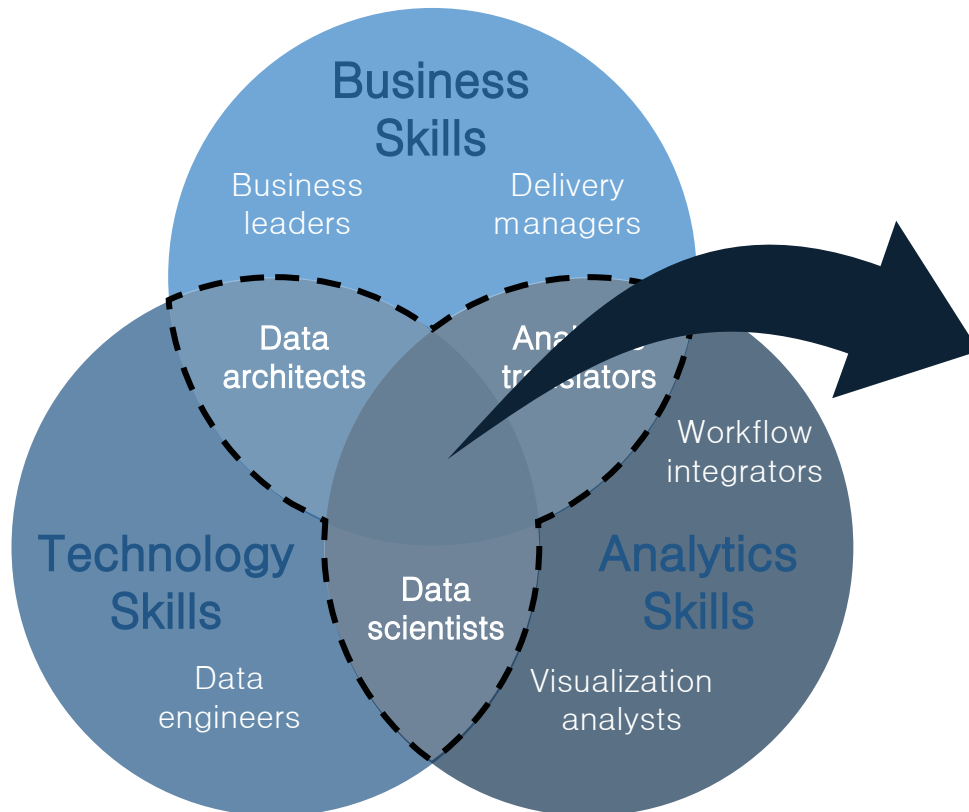
# Index

1. 모의경진대회 개괄
2. AI CONNECT 플랫폼 소개
3. 과제 - Default Prediction
4. 평가지표 - Macro F1 Score
5. 모델 - XGBoost & Ensemble



# MINDS AND COMPANY

AI 기술과 비즈니스의 간극을 좁히는 AI Translator, MNC  
빅데이터, 인공지능 기술에 대한 이해를 바탕으로 최적의 디지털 혁신 전략을 제공하는  
전문 컨설팅 기업



## MNC 지향점

비즈니스, 분석, IT 지식을  
복합적으로 이해하고  
실행하는 Hybrid 역량

# MINDS AND COMPANY

## MNC 사업 영역

MINDs®company

### AI / 데이터 기반 혁신 전략 수립

- **AI/데이터 기반 운영 전략**
  - AI/데이터 기반 운영 기획
  - AI/데이터 기반 프로세스 최적화
- **Tech Sensing**
  - AI Tech. Map 기반 체계적 분류
  - State-of-the-Art 기반 기술 모니터링
- **조직 AI 진단**
  - 기업의 AI 역량 진단
  - AI 혁신 우선순위 과제 도출
- **PoC 기반 혁신 실행 지원**
  - 실제 업무 현장 수준 문제 정의
  - 빠른 성공 사례 구축으로 혁신 확산

### 머신러닝 모델 개발

- **딥러닝 모델 개발**
  - 로그, 정형 데이터 딥러닝 모델
  - Computer Vision 이미지 분류 및 합성 모델
- **자연어 처리(NLP) 분석 모델 개발**
  - STT/TA 활용 NLP 분석 모델
  - 비정형 데이터 기반 추천 모델
- **AI 기반 Business Intelligence 혁신**

### AI / 데이터 솔루션 개발

- **Customized AI 솔루션 개발**
  - 다양한 AI 모델을 활용해 고객사의 니즈에 맞게 구성하는 Tailored 솔루션 시스템
  - 기존 시스템과 솔루션간 연계
  - 다양한 Vendor 활용한 최적화 (자사 및 제휴사의 다양한 솔루션 활용)
- **AI, 시스템 프로젝트 ISP/PMO**
  - AI 솔루션을 활용하는 최적 비즈니스 프로세스 기획
  - 수행 가능 AI 솔루션 프로젝트
- **AI 전문 솔루션 개발 및 운영**
  - AI Connect : AI 전문 경진대회 솔루션
  - AI ML Ops : AI 모델 운영 솔루션

# MINDS AND COMPANY

2017년말 설립 후 40여개 고객사 프로젝트를 성공적으로 수행

## 금융/보험업



## 통신/IT업



## 제조업



## 정부/공공기관/공기업



# MINDS AND COMPANY



## 중고폰 ATM 민팅

당신의 중고폰  
서랍 속에서 민팅 속으로

중고폰ATM · 중고폰판매 · 중고폰기부



시세조회



위치찾기



리뷰왕 이벤트



Make the  
Most of  
AI,



MINDS AND COMPANY

# AI CONNECT

- MNC에서 21년 5월 론칭한 AI 경진대회 플랫폼
- 총 7개 대회 / 22개 과제 성공적 진행 경험 보유
- 국내 최고 AI 경진대회 플랫폼으로 성장 중 (/^ 3^)/
- 이어드림 모의경진대회 진행 예정

과제 목록

[이미지] 코로나 방역을 위한 마스크 착용 여부 분류

주최 경기도

주관 경기도경제과학진흥원

유형 개방형 | 이미지

[자연어]소상공인 QnA 카테고리 분류

주최 경기도

주관 경기도경제과학진흥원

유형 개방형 | 자연어

ai connect.

경진대회

과제

News

로그인

No.1 AI Competition Platform

인공지능 문제해결에 참여해보세요.

AI 전문 문제출제

맞춤형 운영

편리한 기능

철저한 보안

홈페이지 접속하기

AI CONNECT Main Page : <https://main.aiconnect.kr>

## Single Sign On

YEAR-DREAM  
SCHOOL

Click!

강 의 실	채용공고	AI학습	AI Connect	복습과정	학습커뮤니티
 <p>글로벌창업사관학교 부설 스타트업 청년인재 이어드림 스쿨</p>					



## Single Sign On

- 이어드림 스쿨페이지 (<https://yeardreamschool.hunet.co.kr/Home>) 를 통해 AI CONNECT 플랫폼으로 넘어가면, 로그인 및 대회 참여까지 자동 완료
- 첨부 이미지의 '과제 참여 중'이라고 쓰인 부분이 수강생에게는 '과제 참여' 버튼으로 보일텐데, 과제 참여 버튼 클릭시 과제 참여까지 완료됨

### 이어드림 모의경진대회 개인전

중소벤처기업진흥공단 | 마인즈엔컴퍼니 | 모의경진대회

총 상금 원



과제 참여  
버튼 Click!



일정 2022.01.18 ~ 2022.02.08 13:00

참여인원 79명

접수마감 D-1


대회 참여중

# 코드 제출 안내

## 이어드림 모의경진대회 개인전

중소벤처기업진흥공단 | 마인즈엔컴퍼니 | 모의경진대회

총 상금 원



일정 2022.01.18 ~ 2022.02.08 13:00

참여인원 80명

접수마감 D-1

대회 참여중

대회개요

과제개요

데이터

코드공유

리더보드

결과제출

일정


과제문의

팀

규정

Step 1

파일 업로드



파일을 드래그하거나 선택하세요.




- Task Prediction 결과는 csv 파일로 플랫폼에 제출
- 최종 제출 및 과제 종료 후 추론에 쓰인 코드 파일(.ipynb)은 기존 일일 과제 제출 포맷과 유사한 형태로 제출 예정(코드 파일 제출 방법은 추후 공지)



# Index

1. 모의경진대회 개괄
2. AI CONNECT 플랫폼 소개
3. 과제 - Default Prediction
4. 평가지표 - Macro F1 Score
5. 모델 - XGBoost & Ensemble

# 모의경진대회 과제 개괄

개인전	팀전																																						
대부업체 고객 데이터 통한 채무 불이행 예측	흉부 CT 이미지 통한 COVID 감염 예측	고속도로 교통량 예측	악성 댓글 분류																																				
<table border="1"> <thead> <tr> <th></th><th>int_rate</th><th>annual_inc</th><th>dti</th></tr> </thead> <tbody> <tr> <td>count</td><td>100000.000000</td><td>1.000000e+05</td><td>100000.000000</td></tr> <tr> <td>mean</td><td>0.130833</td><td>7.436061e+04</td><td>18.514508</td></tr> <tr> <td>std</td><td>0.044773</td><td>7.467409e+04</td><td>8.413049</td></tr> <tr> <td>min</td><td>0.053200</td><td>5.360000e+03</td><td>0.000000</td></tr> <tr> <td>25%</td><td>0.097500</td><td>4.500000e+04</td><td>12.200000</td></tr> <tr> <td>50%</td><td>0.127400</td><td>6.200000e+04</td><td>18.060000</td></tr> <tr> <td>75%</td><td>0.158000</td><td>9.000000e+04</td><td>24.530000</td></tr> <tr> <td>max</td><td>0.309900</td><td>8.300000e+06</td><td>49.930000</td></tr> </tbody> </table>		int_rate	annual_inc	dti	count	100000.000000	1.000000e+05	100000.000000	mean	0.130833	7.436061e+04	18.514508	std	0.044773	7.467409e+04	8.413049	min	0.053200	5.360000e+03	0.000000	25%	0.097500	4.500000e+04	12.200000	50%	0.127400	6.200000e+04	18.060000	75%	0.158000	9.000000e+04	24.530000	max	0.309900	8.300000e+06	49.930000			
	int_rate	annual_inc	dti																																				
count	100000.000000	1.000000e+05	100000.000000																																				
mean	0.130833	7.436061e+04	18.514508																																				
std	0.044773	7.467409e+04	8.413049																																				
min	0.053200	5.360000e+03	0.000000																																				
25%	0.097500	4.500000e+04	12.200000																																				
50%	0.127400	6.200000e+04	18.060000																																				
75%	0.158000	9.000000e+04	24.530000																																				
max	0.309900	8.300000e+06	49.930000																																				
<b>Task</b> : 금융 & 정형 시작 : 01/26 09:00 AM 종료 : 02/08 12:00 PM	<b>Task</b> : 바이오 & 이미지 시작 : 02/09 09:00 AM 종료 : 추후 공지	<b>Task</b> : 유통 & 정형 시작 : 02/16 09:00 AM 종료 : 추후 공지	<b>Task</b> : 게임 & 자연어 시작 : 02/23 09:00 AM 종료 : 추후 공지																																				
개인전 성적 통한 팀 매칭	팀별 GPU 제공	팀별 GPU 제공	팀별 GPU 제공																																				

# Default Prediction

## 대부업체 고객 데이터를 통한 채무 불이행 예측 모델 | 정형 & 금융

대부업체 고객 데이터를 통해 고객 별 채무 불이행 여부를 이진 분류하는 과제

데이터셋

### • 데이터 구조

- train.csv (100000 rows X 76 columns) : 75개의 feature 및 depvar(dependent variable)
- test.csv (35816 rows X 76 columns) : 75개의 feature 및 ID
- sample\_submission.csv (35816 rows X 2 columns) : answer 및 ID

변수

### • 변수 구성

- Input : 대부업체의 고객 정보를 나타내는 76개의 변수
- Output : 채무 불이행 여부를 뜻하는 depvar(train.csv) or answer(test.csv 및 sample\_submission.csv)
- ID : 성능 평가를 위한 인덱스

	int_rate	annual_inc	dti	delinq_2yrs	inq_last_6mths	pub_rec	revol_bal	total_acc	collections_12_mths_ex_med	acc_now_delinq	...
0	0.0824	21000.0	29.19	0	1	0	3016	26	0	0	...
1	0.1299	80000.0	4.82	0	1	1	5722	24	0	0	...
2	0.1299	38000.0	23.66	0	3	0	6511	18	0	0	...
3	0.1367	100000.0	16.27	4	2	0	6849	30	0	0	...

train.csv

# Default Prediction


데이터 다운로드 링크 : [LINK](#)


공유 문서함 > 중진공-이어드림 프로젝트 교육 운영 > 03. 강의 자료 및 영상 > 28. 20220126\_DAY27(1주차 모의경진대회) ▾





이름 ↑	소유자	마지막으로 수정한 날짜	파일 크기
 sample_submission.csv 	나	오후 10:19 나	269KB
 test.csv 	나	오후 10:19 나	7.4MB
 train.csv 	나	오후 10:19 나	20.4MB




수강생 공유 폴더함에서 다운로드하여 개인 드라이브에 업로드


 드라이브











 새로 만들기


☐ 우선순위


 내 드라이브

 공유 드라이브


 공유 문서함


 최근 문서함


 중요 문서함

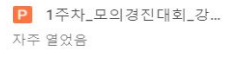
 휴지통

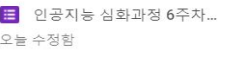
내 드라이브 ▾




 train.csv  
방금 업로드함

 test.csv  
방금 업로드함

 sample\_submission.csv  
방금 업로드함





이름 ↑	소유자	마지막으로 수정한 날짜	파일 크기
 sample_submission.csv	나	오전 5:19 나	269KB
 test.csv	나	오전 5:19 나	7.4MB
 train.csv	나	오전 5:19 나	20.4MB

# Index

1. 모의경진대회 개괄
2. AI CONNECT 플랫폼 소개
3. 과제 - Default Prediction
4. 평가지표 - Macro F1 Score
5. 모델 - XGBoost & Ensemble

## 정확도(Accuracy)의 한계

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

- 다음과 같은 2-class 문제를 고려해보자.
  - NO class 샘플 수 = 990
  - YES class 샘플 수 = 10
- 모델이 모든 샘플에 대해 NO를 예측하기만 하더라도, 정확도(accuracy)=99% 달성
  - 이 모델은 단 하나의 YES도 예측하지 않음
  - 보통 더 귀한 클래스를 잘 예측하는 것이 중요 (ex) frauds, intrusions, defects)
- Imbalanced Data의 경우에는 대안적 평가 지표 모색 필요

# Precision and Recall

$$\text{F1-Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

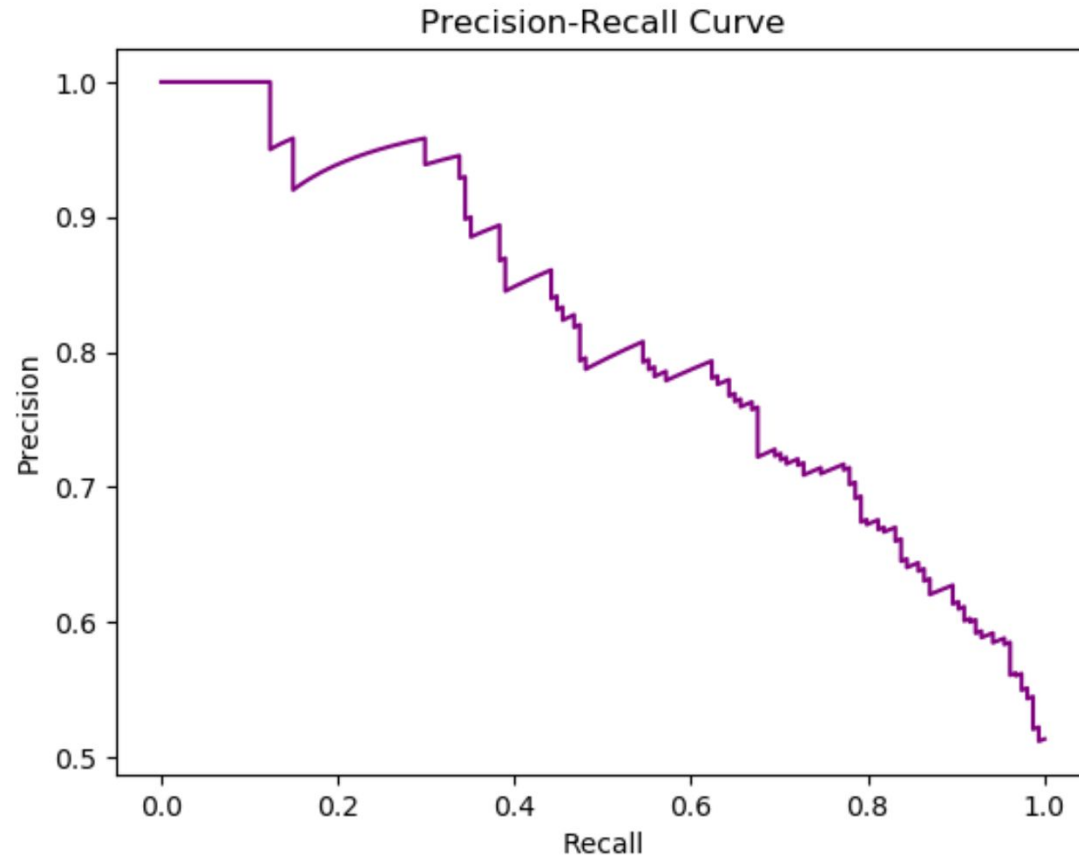
$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

# PR Curve



- PR(Precision - Recall Curve)를 그려보면 우하향하는 형태가 일반적
- Precision - Recall 간 trade-off 관계 확인 가능



## Micro F1 Score

		Predicted		
		Cat	Fish	Hen
True	Cat	4	1	1
	Fish	6	2	2
	Hen	3	0	6

**Micro-averaging** : 각 샘플에 동일한 가중치를 적용하여 평균하는 방법

$$\begin{aligned}
 \text{Micro - precision} &: \frac{TP_{Cat} + TP_{Fish} + TP_{Hen}}{TP_{Cat} + TP_{Fish} + TP_{Hen} + FP_{Cat} + FP_{Fish} + FP_{Hen}} \\
 &= \frac{4 + 2 + 6}{4 + 2 + 6 + (6 + 3) + (1 + 0) + (1 + 2)} = \frac{12}{12 + 13} = 0.48
 \end{aligned}$$

$$\begin{aligned}
 \text{Micro - recall} &: \frac{TP_{Cat} + TP_{Fish} + TP_{Hen}}{TP_{Cat} + TP_{Fish} + TP_{Hen} + FN_{Cat} + FN_{Fish} + FN_{Hen}} \\
 &= \frac{4 + 2 + 6}{4 + 2 + 6 + (1 + 1) + (6 + 2) + (3 + 0)} = \frac{12}{12 + 13} = 0.48
 \end{aligned}$$

# Macro F1 Score

		Predicted		
True		Cat	Fish	Hen
	Cat	4	1	1
	Fish	6	2	2
	Hen	3	0	6

Class-wise  
result

Class	Precision	Recall	F1-score
Cat	30.8%	66.7%	42.1%
Fish	66.7%	20.0%	30.8%
Hen	66.7%	66.7%	66.7%

**Macro-averaging** : 각 클래스에 동일한 가중치를 적용하여 평균하는 방법

$$\text{F1-score(Cat)} = 2 \times (30.8\% \times 66.7\%) / (30.8\% + 66.7\%) = 42.1\%$$

$$\text{Macro-F1} = (42.1\% + 30.8\% + 66.7\%) / 3 = 46.5\%$$

## Weighted F1 Score

Class	Precision	Recall	F1-score
Cat	30.8%	66.7%	42.1%
Fish	66.7%	20.0%	30.8%
Hen	66.7%	66.7%	66.7%

	Cat	Fish	Hen
Cat	4	1	1
Fish	6	2	2
Hen	3	0	6

**Weighted-averaging** : 실제 인스턴스 수에 따라 클래스의 점수에 가중치를 부여하여 계산하는 방법  
클래스의 불균형을 다룰 때 유용

$$\text{Weighted-F1} = (6 \times 42.1\% + 10 \times 30.8\% + 9 \times 66.7\%) / 25 = 46.4\%$$

# Index

1. 모의경진대회 개괄
2. AI CONNECT 플랫폼 소개
3. 과제 - Default Prediction
4. 평가지표 - Macro F1 Score
5. 모델 - XGBoost & Ensemble

## Ensemble이란?

참고자료 : [LINK](#)

**앙상블(ensemble)**이란, 쉽게 말하면, 비슷한 무리들의 집합

**앙상블(ensemble)**이란, 여러 모델들에서 나온 결과들에 대해,  
평균치를 내거나 다수결을 위한 투표를 하는 등  
여러 모델들의 집단 지성을 활용하여  
단일 모델보다 더 나은 결과를 도출해 내는 방법

앙상블(집단 지성을 활용하는 방법)에도 다양한 방법이 있다.

**Voting** - 투표를 통해 결과 도출

**Bagging** - Bootstrap Aggregating (복원 추출로 다양한 샘플 생성)

**Boosting** - 이전 오차를 보완하며 가중치 부여

**Stacking** - 여러 모델이 예측한 결과를 다시 학습

앙상블 기법에는 다양한 방식들이 추가로 더 있을 수 있지만,  
위에 언급한 4가지가 대표적인 앙상블 기법이며,  
sklearn에 이미 잘 구현되어 있다.



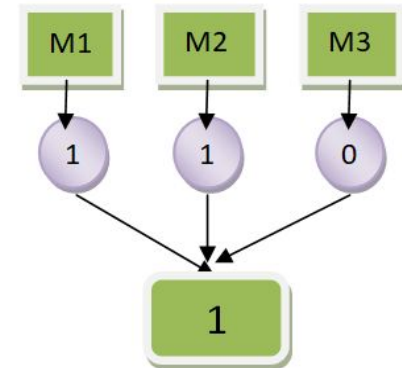
# Ensemble - Voting

참고자료 : [LINK](#)

**Voting**은 투표를 통해 결정하는 방식

## Hard Voting Classifier

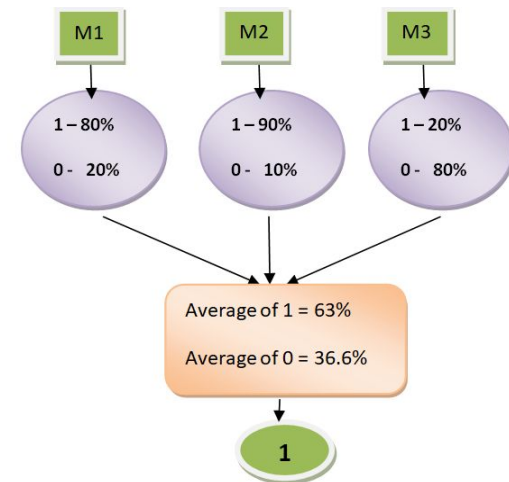
여러 모델을 생성하고 그 결과를 비교.  
분류기들의 결과들을 집계하여,  
가장 많은 표를 얻은 클래스를 최종 예측값으로 결정



Hard Voting

## Soft Voting Classifier

앙상블에 사용되는 모든 분류기가 클래스의 확률을  
예측할 수 있을 때 사용.  
분류기들의 클래스별 예측 확률을 평균하여,  
확률이 가장 높은 클래스를 최종 예측값으로 결정



Soft Voting

보통 대회에서는 Hard Vote 보다는 Soft Vote 선호

# Ensemble - Bagging

참고자료 : [LINK](#)

## - Bootstrap(부트스트랩)

주어진 dataset로부터 복원추출을 통해 여러 sample dataset을 생성하는 방법

## - Bagging(배깅)

부트스트랩을 통해 생성한 각각의 sample dataset에 개별 모델을 적용하고, 그 결과들을 종합

**Classification** : 개별 모델 결과들에 대한 투표(Voting)

**Regression** : 개별 모델 결과들에 대한 평균

## - Voting과 Bagging의 차이점

**Voting** : 다른 알고리즘 모델들이 도출한 결과에 대한 투표

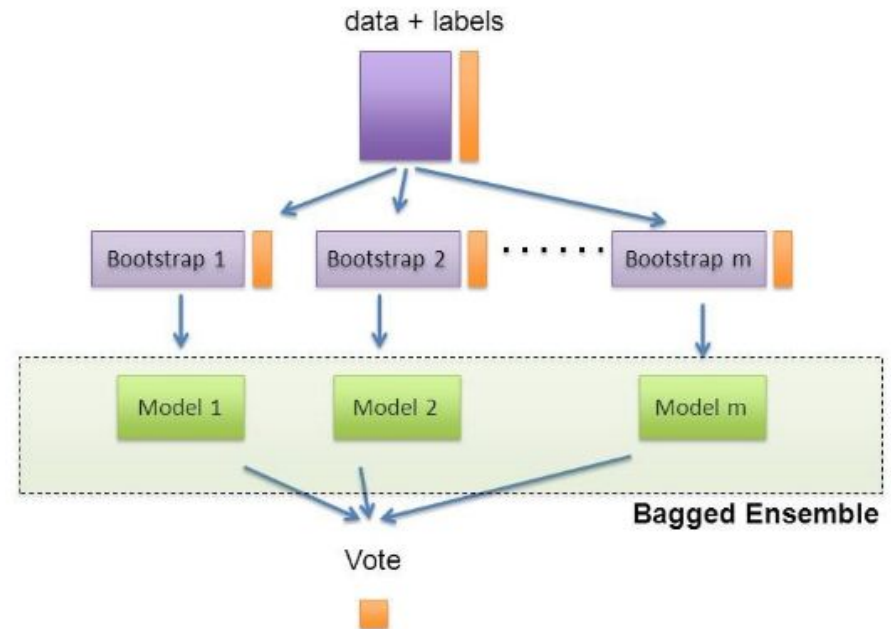
**Bagging** : 다른 sample dataset에 같은 알고리즘 모델을 적용하여 도출한 결과에 대한 투표

## - RandomForest

대표적인 Bagging 방식의 알고리즘

다른 sample dataset에 동일 모델(Decision Tree) 적용

“Bagging” : Bootstrap **AGG**regating



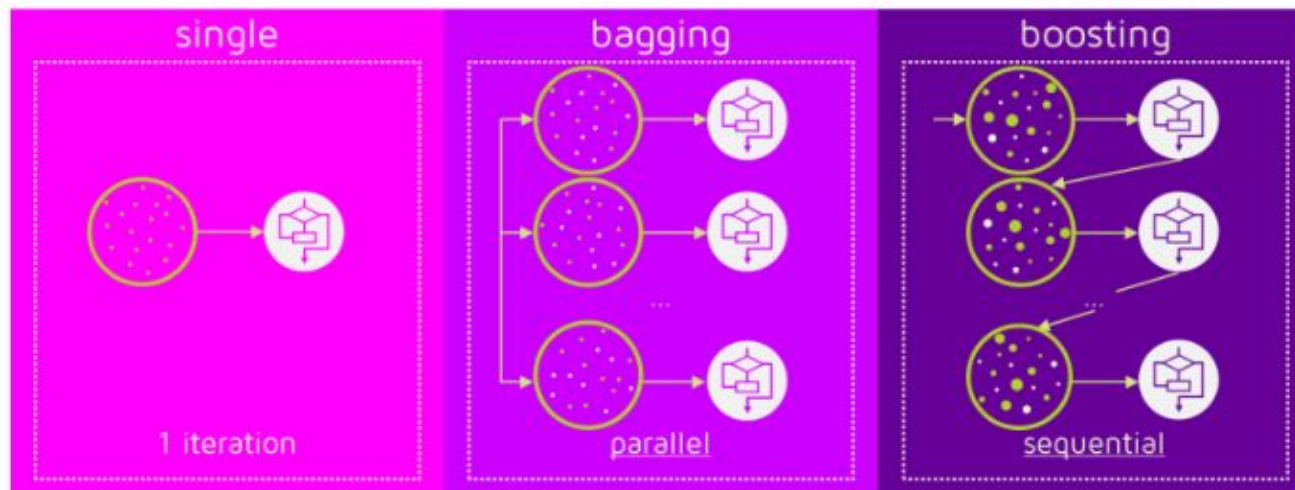


# Ensemble - Boosting

참고자료 : [LINK](#)

Boosting은 앙상블 학습(ensemble learning)의 하나로서, 약한 학습기를 **순차적으로 학습**하되, 이전 학습에서 **잘못 예측된 데이터에 더 큰 가중치를 부여**해 오차를 보완해 나가는 알고리즘. 순차적이기 때문에 병렬 처리에 어려움이 있고, 그렇기 때문에 다른 앙상블 대비 학습 시간이 오래걸린다는 단점 존재.

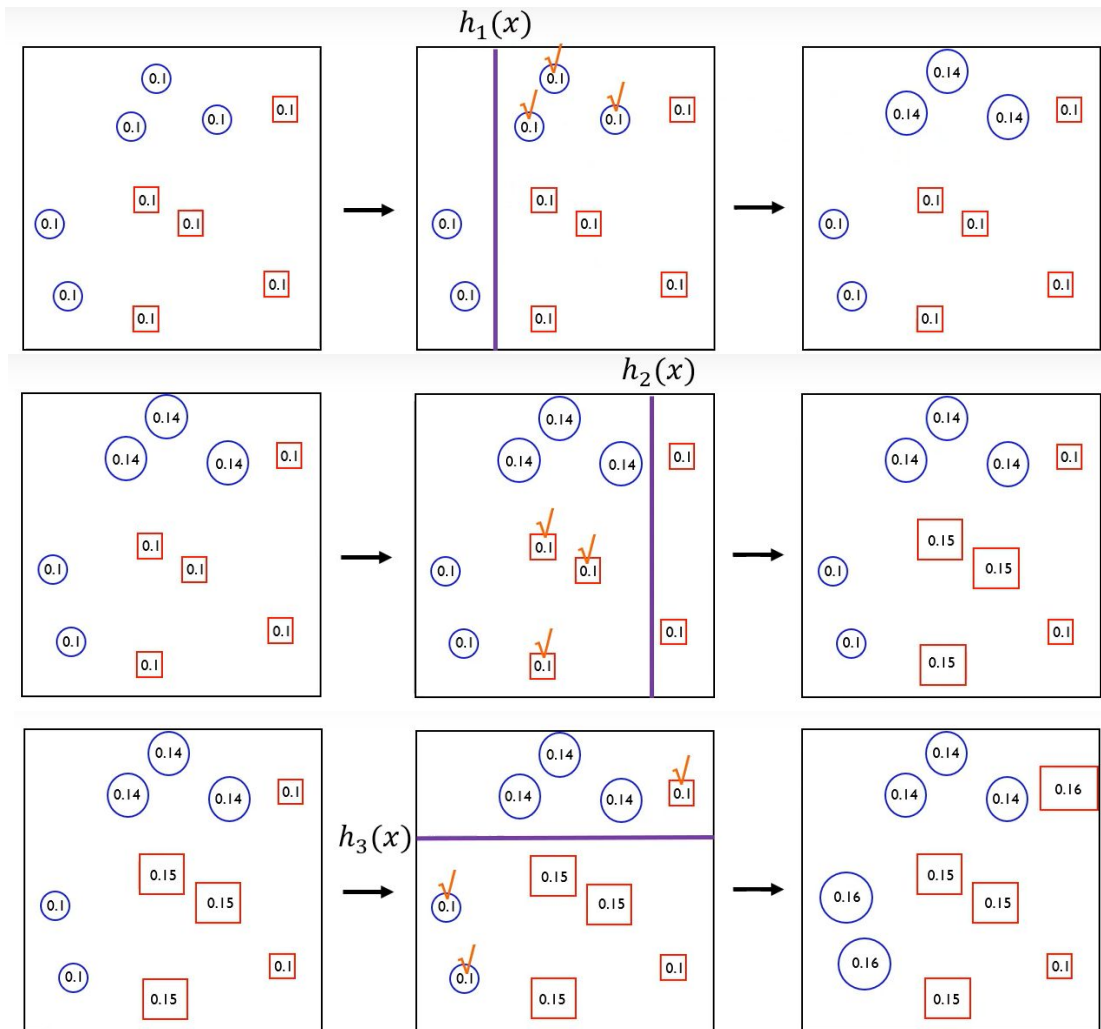
Boosting 알고리즘의 예로는 1. **Adaboost**(Adaptive boosting), 2. **GBM**(Gradient boosting machines), 3. **XGBoost**(eXtreme Gradient Boosting), 4. **LGBM**(Light gradient boost machines), 5. **CATBoost** 등이 있다.





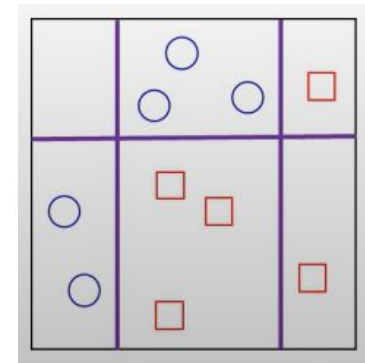
# Ensemble - Boosting - AdaBoost

참고자료 : [LINK](#)

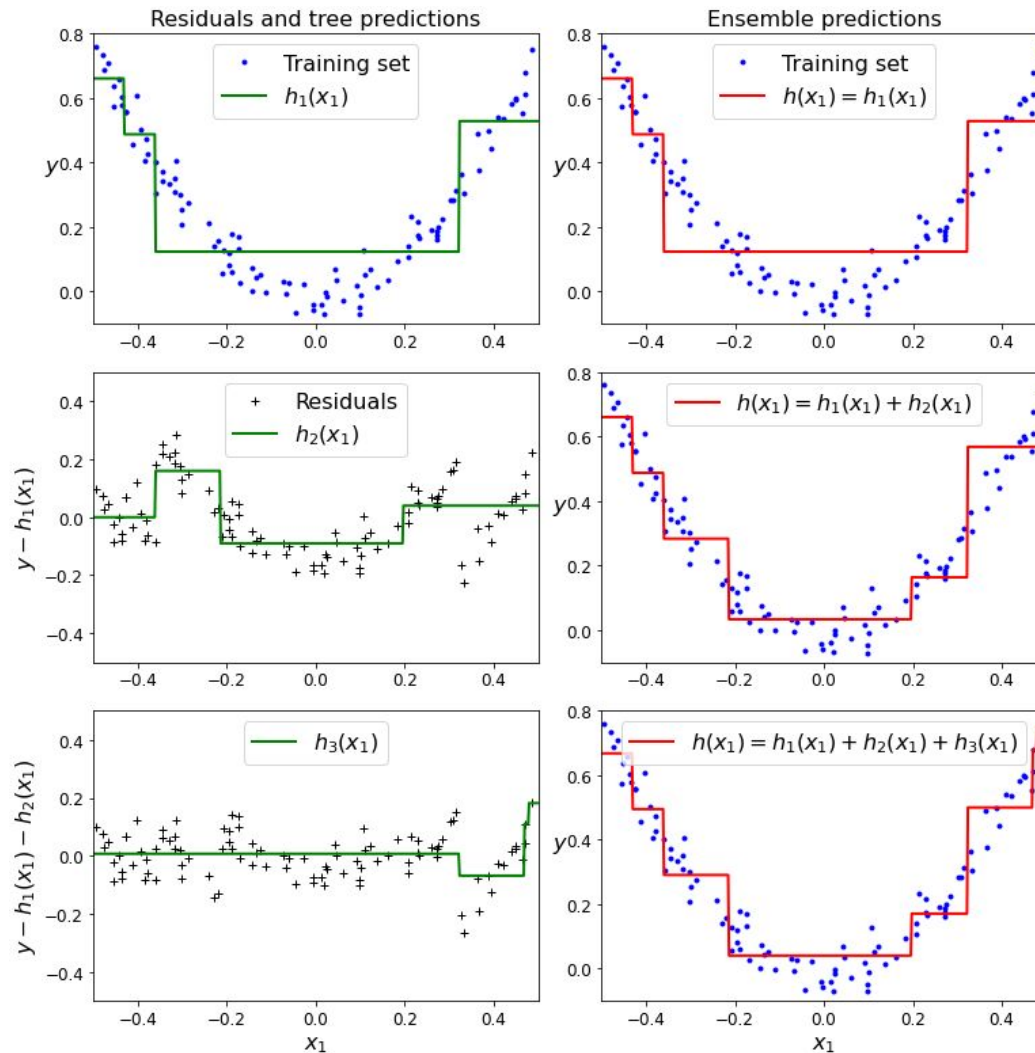


에이다부스트(AdaBoost)는 Adaptive Boost의 줄임말로 이전 예측기가 잘 맞추지 못했던 샘플의 가중치를 높임으로써 이전 학습기를 보완해 나가는 방법

$$h(x) =$$



# Ensemble - Gradient Boosting

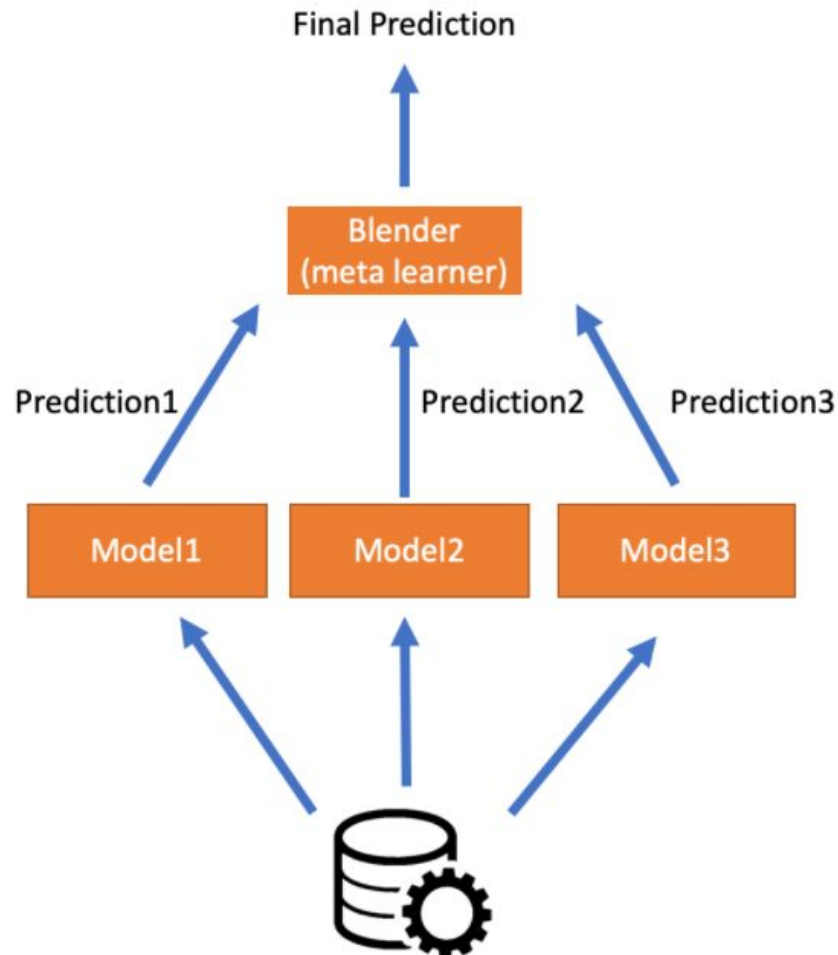


그레이디언트 부스팅은  
에이다부스트처럼 이전까지의 오차를  
보정하도록 예측기를 순차적으로 추가

하지만 에이다부스트처럼 반복마다  
샘플의 가중치를 수정하는 대신 이전  
예측기가 만든 잔여오차 (**residual  
error**)에 새로운 예측기를 학습시킴

# Ensemble - Stacking

참고자료 : [LINK1](#), [LINK2](#)



스태킹(Stacking)은 stacked generalization의 줄임말

여러 개의 개별 알고리즘을 합쳐 예측을 한 후, 개별 알고리즘으로 예측한 데이터를 기반으로 다시 예측 과정 수행



# End of document

Contact: (주)마인즈앤컴퍼니 고석태 대표 (stko@mnc.ai)