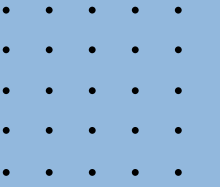


# 메타모델 기반 앙상블 기법 : 스택킹과 블렌딩의 성능 비교

AI 엔지니어링 | 김지윤



# 목차



## 1. 주제 선정 이유

## 2. 메타모델이란?

1. 앙상블
2. 메타모델
3. 스택킹, 블렌딩

## 3. 실습

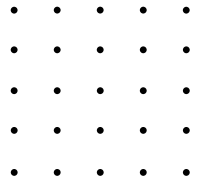
1. 사용한 데이터
2. 구현 과정
3. 모델 성능 비교

## 4. 결론

1. 그래서 무엇을 선택해야 할까?
2. 이슈 / 발전방향

# 1. 주제 선정 이유

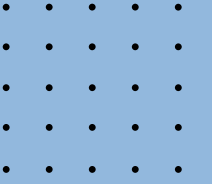
우리 FISA 5기  
2차 기술 세미나



## 주제 선정 이유

앙상블 기법을 직접 실습하여 스택킹과 블렌딩의 성능 차이를 비교하여  
머신러닝 성능을 향상시킬 수 있는 방법을 알아보기 위함





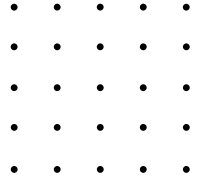
# 메타모델이란?

- 앙상블
- 메타모델
- 스택킹, 블렌딩



## 2. 메타모델이란? | 앙상블

우리 FISA 5기  
2차 기술 세미나



# 앙상블

여러 머신러닝 모델을 결합하여 각각의 모델에서 발생할 수 있는 오류를 줄이고  
전체적인 예측 성능을 향상시키는 기법

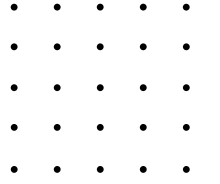
## 왜 앙상블 사용?

- 성능향상
- 과적합 방지



## 2. 메타모델이란? | 앙상블

우리 FISA 5기  
2차 기술 세미나



### 앙상블 기법 종류

#### 배깅 (Bagging)

여러 버전의 훈련 데이터로  
병렬적으로 모델을 만든 후  
결과를 합치는 방식  
(랜덤 포레스트)

#### 부스팅 (Boosting)

같은 유형의 모델을  
순차적으로 학습시키며  
이전 모델의 오류를 보완해나가는 방식  
(XGBoost, LightGBM)

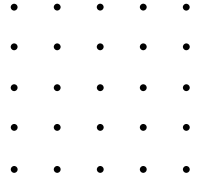
#### 스태킹 (Stacking)

여러 다른 종류의 모델들의  
예측값을 입력으로 받아  
이를 학습하는 새로운 메타 모델을  
만들어 최종 예측을 하는 방식  
(기본모델+메타모델)



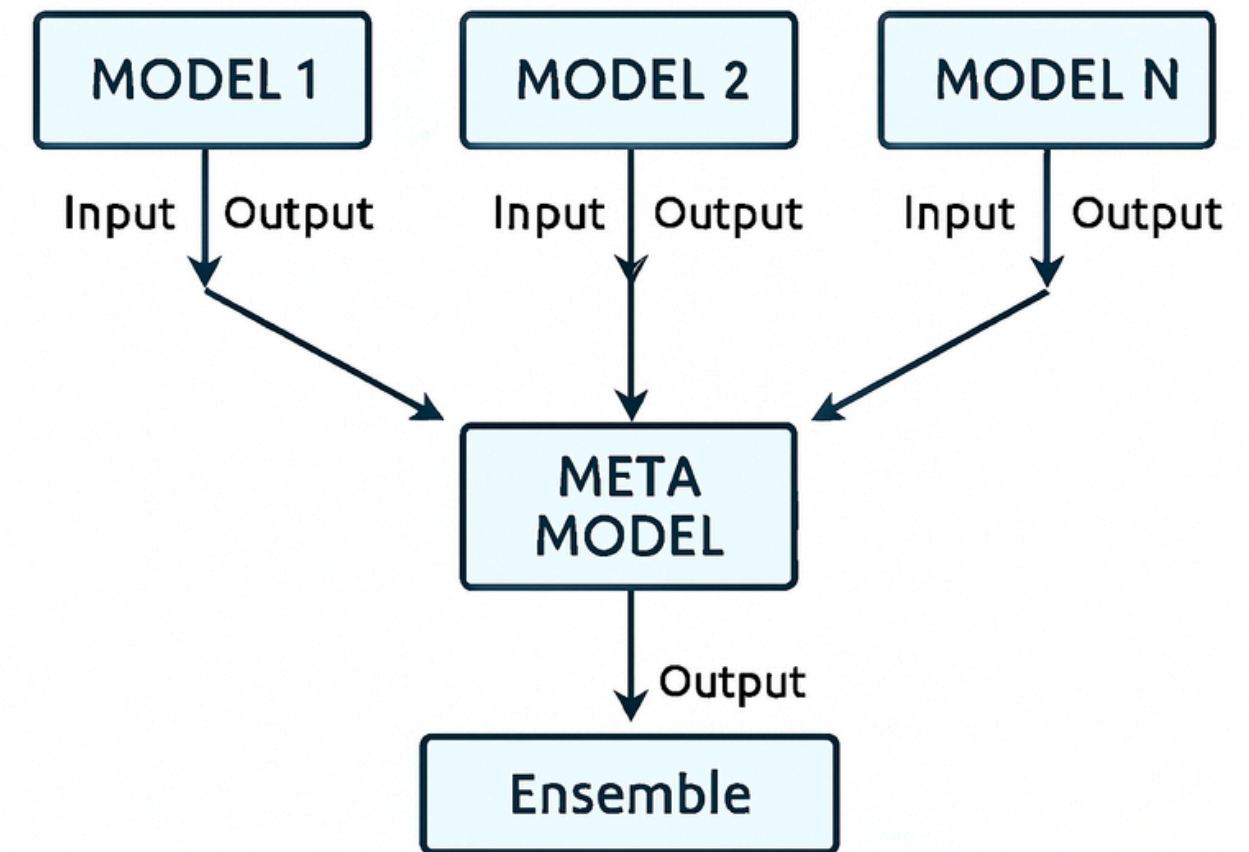
## 2. 메타모델이란? | 메타모델

우리 FISA 5기  
2차 기술 세미나



# 메타모델

모델을 학습하는 모델  
즉, 다른 모델의 예측값을 학습하는 모델

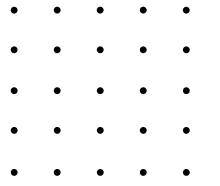


[사진1] 메타모델 아키텍처 다이어그램



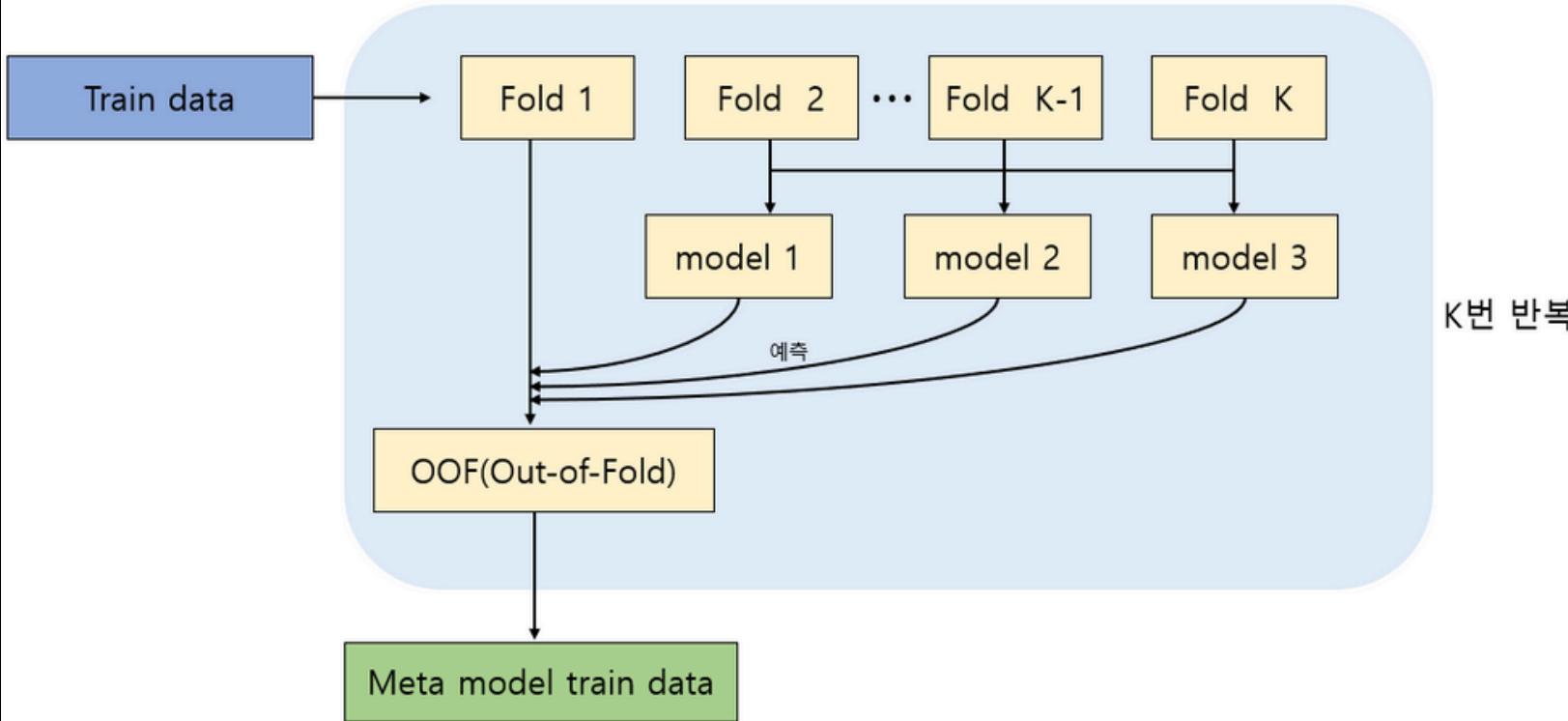
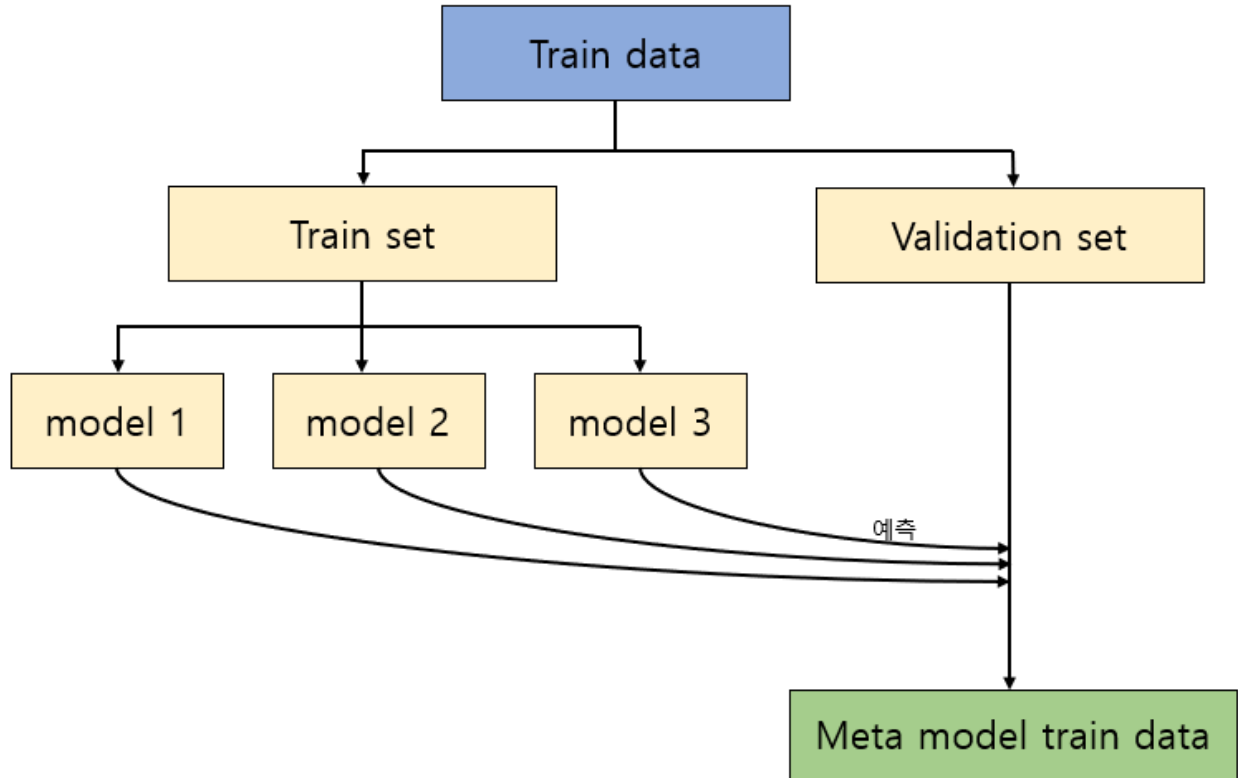
## 2. 메타모델이란? | 스택킹, 블렌딩

우리 FISA 5기  
2차 기술 세미나



# 스택킹 / 블렌딩

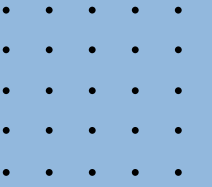
“메타모델을 학습시키기 위한 데이터를 어떻게 만드는가”

| 스택킹(Stacking)   | 블렌딩(Blending)   |
|---|---|
| K-fold 교차 검증  | 홀드아웃 분할(Hold-out)   |
|  <p>The diagram illustrates the Stacking process using K-fold cross-validation. It starts with 'Train data' (blue box) which is split into 'Fold 1', 'Fold 2', ..., 'Fold K-1', and 'Fold K' (yellow boxes). For each fold, a model (model 1, model 2, model 3) is trained on the training data and predicts the out-of-fold data. These predictions are used as input for the meta-model training data (green box). The process is repeated K times, as indicated by 'K번 반복' (K times repeat).</p> |  <p>The diagram illustrates the Blending process using hold-out split. It starts with 'Train data' (blue box) which is split into 'Train set' and 'Validation set' (yellow boxes). The 'Train set' is used to train multiple models (model 1, model 2, model 3) (yellow boxes). The 'Validation set' is used to evaluate the models and generate predictions, which are then used as input for the meta-model training data (green box).</p> |

[그림2] 스택킹 기반 메타모델 생성 과정

[그림3] 블렌딩 기반 메타모델 생성 과정





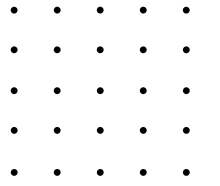
# 실습

- 데이터 설명
- 구현 과정
- 모델 성능 비교



# 3. 실습 | 데이터 설명

우리 FISA 5기  
2차 기술 세미나



## Kaggle의 “사기 거래 탐지” 데이터

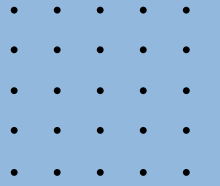
|   | Unnamed: 0 | TRANSACTION_ID | TX_DATETIME         | CUSTOMER_ID | TERMINAL_ID | TX_AMOUNT | TX_TIME_SECONDS | TX_TIME_DAYS | TX_FRAUD | TX_FRAUD_SCENARIO |
|---|------------|----------------|---------------------|-------------|-------------|-----------|-----------------|--------------|----------|-------------------|
| 0 | 0          | 0              | 2023-01-01 00:00:31 | 596         | 3156        | 533.07    | 31              | 0            | 0        | 0                 |
| 1 | 1          | 1              | 2023-01-01 00:02:10 | 4961        | 3412        | 808.56    | 130             | 0            | 0        | 0                 |
| 2 | 2          | 2              | 2023-01-01 00:07:56 | 2           | 1365        | 1442.94   | 476             | 0            | 1        | 1                 |
| 3 | 3          | 3              | 2023-01-01 00:09:29 | 4128        | 8737        | 620.65    | 569             | 0            | 0        | 0                 |
| 4 | 4          | 4              | 2023-01-01 00:10:34 | 927         | 9906        | 490.66    | 634             | 0            | 0        | 0                 |

|                |   |                   |   |
|----------------|---|-------------------|---|
| Unnamed: 0     | 0 | TX_AMOUNT         | 0 |
| TRANSACTION_ID | 0 | TX_TIME_SECONDS   | 0 |
| TX_DATETIME    | 0 | TX_TIME_DAYS      | 0 |
| CUSTOMER_ID    | 0 | TX_FRAUD          | 0 |
| TERMINAL_ID    | 0 | TX_FRAUD_SCENARIO | 0 |

| TX_FRAUD   |          |
|------------|----------|
| 0: 합법적인 거래 | 1: 사기 거래 |
| 86.50%     | 13.50%   |

# 3. 실습 | 구현 과정

우리 FISA 5기  
2차 기술 세미나



## 사용모델 및 평가지표

기본  
모델

LightGBM, RandomForest

메타  
모델

Logistic Regression

성능  
평가

f1-score\*, PR-AUC(AUPRC)\*\* , ROC-AUC\*\*\*,

\*f1-score: 정밀도(Precision)와 재현율(Recall)의 조화평균으로, 특히 데이터가 불균형할 때 모델의 실질적인 성능을 잘 보여주는 핵심 평가지표 중 하나

\*\*PR-AUC(AUPRC): 이진 분류 모델의 성능을 평가하는 지표 중 하나로, Positive(소수) 클래스를 얼마나 정밀하게 빠짐없이 잘 찾아내는지

\*\*\*ROC-AUC: Positive 클래스와 Negative 클래스를 얼마나 잘 구별하는지

# 3. 실습 | 구현 과정

우리 FISA 5기  
2차 기술 세미나

· · · · ·  
· · · · ·  
· · · · ·  
· · · · ·  
· · · · ·

01

## 스태킹

```
# --- 스태킹 모델 정의 ---
estimators = [
    ('lgbm', LGBMClassifier(random_state=42)),
    ('rf', RandomForestClassifier(n_estimators=50,
                                random_state=42, n_jobs=-1))
]
final_estimator = LogisticRegression()

stack_model = StackingClassifier(
    estimators=estimators,
    final_estimator=final_estimator,
    cv=5, ← CV=5
    n_jobs=-1
)
```

5개의 fold로 나누어 교차 검증 방식으로  
모델 학습, 예측값 생성

02

## 블렌딩

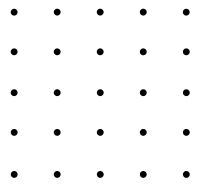
```
# --- 블렌딩 모델 정의 ---
estimators = [
    ('lgbm', LGBMClassifier(random_state=42)),
    ('rf', RandomForestClassifier(n_estimators=50,
                                random_state=42, n_jobs=-1))
]
final_estimator = LogisticRegression()

blend_model = BlendingClassifier(
    estimators=estimators,
    final_estimator=final_estimator,
    test_size=0.5, ← test_size = 0.5
    random_state=42
)
```

홀드아웃 방식으로 데이터를 분리하여  
모델 학습, 예측값 생성

### 3. 실습 | 모델 성능 비교

우리 FISA 5기  
2차 기술 세미나



01

#### 스태킹

스태킹 모델 평가 결과:

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.9946    | 1.0000 | 0.9973   | 455456  |
| 1            | 1.0000    | 0.9651 | 0.9822   | 70791   |
| accuracy     |           |        | 0.9953   | 526247  |
| macro avg    | 0.9973    | 0.9825 | 0.9898   | 526247  |
| weighted avg | 0.9953    | 0.9953 | 0.9953   | 526247  |

ROC-AUC Score: 0.9841  
AUPRC: 0.9764

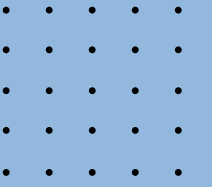
02

#### 블렌딩

블렌딩 모델 평가 결과:

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.9946    | 0.9999 | 0.9973   | 455456  |
| 1            | 0.9994    | 0.9651 | 0.9820   | 70791   |
| accuracy     |           |        | 0.9952   | 526247  |
| macro avg    | 0.9970    | 0.9825 | 0.9896   | 526247  |
| weighted avg | 0.9953    | 0.9952 | 0.9952   | 526247  |

ROC-AUC Score: 0.9835  
AUPRC: 0.9762



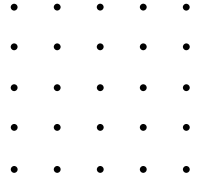
## 결론

- 그래서 무엇을 선택해야 할까?
- 이슈 / 발전방향

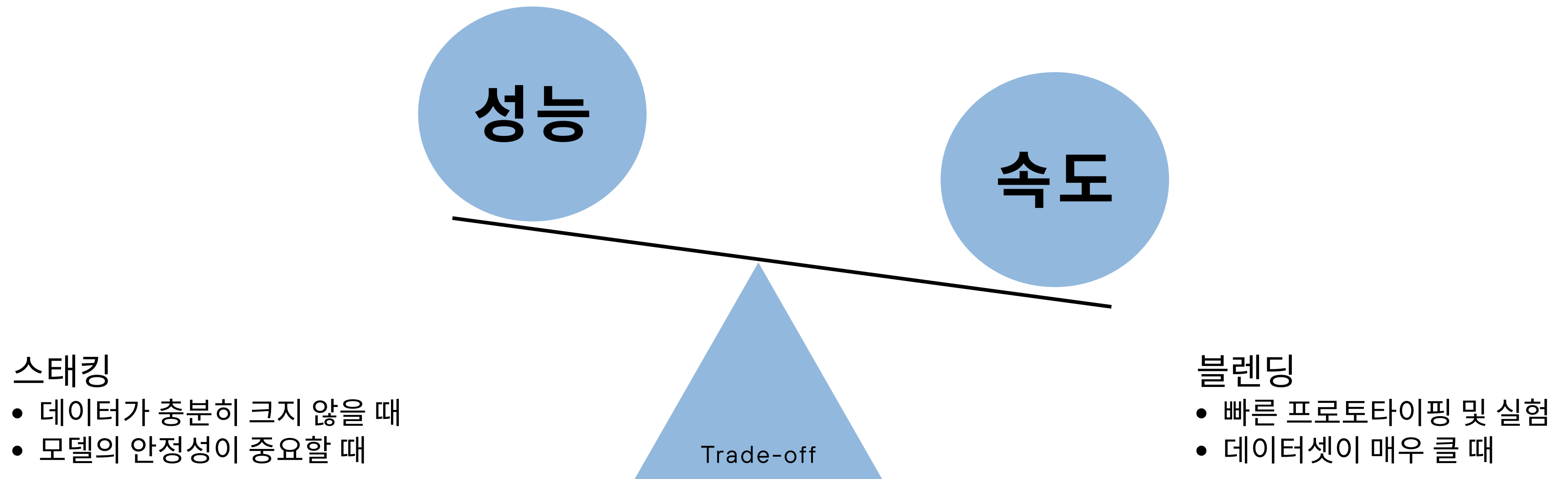


## 4. 결론 | 그래서 무엇을 선택해야 할까?

우리 FISA 5기  
2차 기술 세미나

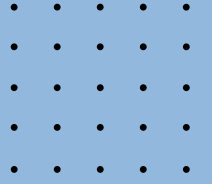


# 언제, 어떤 기법을 사용할까?



## 4. 결론 | 이슈/발전방향

우리 FISA 5기  
2차 기술 세미나



# 이슈

01

### 하이퍼파라미터 최적화 부재:

각 기본 모델에 대한 하이퍼파라미터 튜닝 진행 X

→ 모델들이 가진 최상의 성능을 이끌어내지 못했을 가능성

02

### 제한적인 모델 다양성:

LightGBM, RandomForest 트리 기반 모델 위주로 앙상블을 구성,  
더 다양한 종류의 모델을 사용하지 않아 다양성 극대화하지 못함

03

### 모델 성능의 상한선 효과:

기본 모델들의 성능이 높아 두 앙상블 기법 간의 유의미한 성능 차이 관찰 어려움

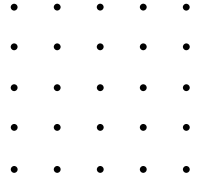
→ 의도적으로 과적합을 유도하여 모델의 안정성 테스트

→ 과적합 상황에서 스테킹이 더 안정적인 성능을 보임



## 4. 결론 | 이슈/발전방향

우리 FISA 5기  
2차 기술 세미나



# 향후 발전 방향

### 하이퍼파라미터 최적화 수행:

PyCaret과 같은 AutoML(자동화 머신러닝) 라이브러리를 통하여

각 기본 모델의 최적의 하이퍼파라미터 탐색

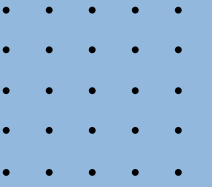
→ 기본 모델의 성능을 최대한으로 끌어올려 앙상블 모델의 전반적인 성능 향상

### 다양한 기본 모델 추가 및 선정:

앙상블에 K-NN, SVM, 신경망(Neural Network) 등

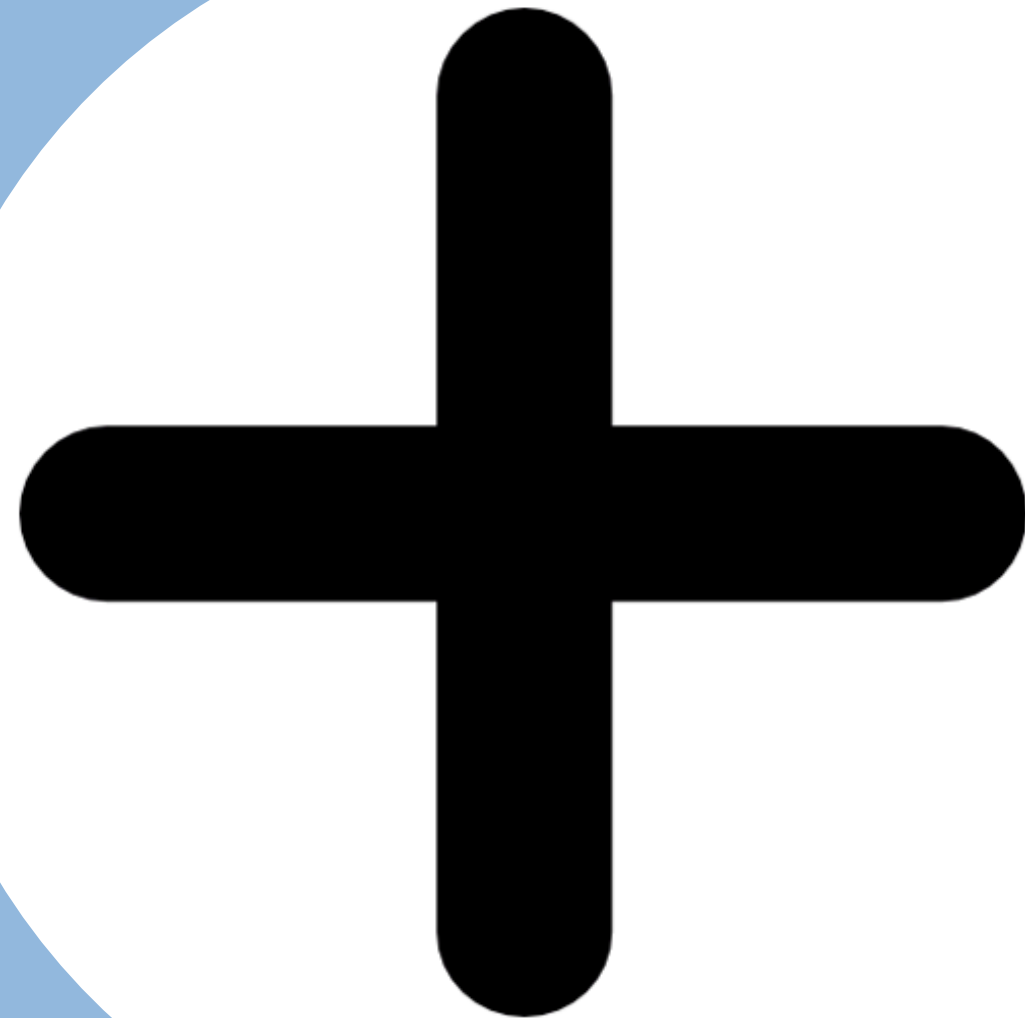
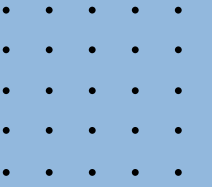
다양한 모델들을 추가하여 다양성 확보,

각 모델의 성능 및 예측 결과의 상관관계를 분석하여 최적의 기본 모델 조합 선정



# 감사합니다.





# 부록

- 실제 금융 활용 사례



2. XGBoost, LightGBM, CatBoost 등을 스택킹 앙상블로 결합해 신용카드 사기 탐지에서 AUC 0.99 이상의 뛰어난 성능을 달성했고, 데이터 불균형 극복과 해석가능성(XAI) 동시에 강화

부록  
실제 금융 활용 사례

AutoML 프레임워크에서 AI모델 성능 고도화를 위한 기술 전략

Stacking ensemble method for personal credit risk assessment in Peer-to-Peer lending

서동준\*, 정경용\*\*

Technical Strategies for Performance Enhancement in AutoML Frameworks

Dongjoon Suh\*, Kyungyong Chung\*\*

이 논문은 2025년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2020-NR049579) 아울러, 본 연구는 2025년 경기대학교 대학원 연구원장학생 장학금

요약

본 논문은 AutoML 프레임워크에서 AI 모델 성능을 향상시키기 위한 기술 전략을 AutoKeras, AutoGluon이라는 세 가지 주요 도구의 구조적 기반, 최적화 접근법, 성능 분석 결과, 도메인 특수성 부족, 분절된 인터페이스, 모델 해석 가능성 부족이라는 문제를 해결하기 위해 도메인 특화 메타러닝, 뉴로심볼릭 통합 인터페이스라는 두 가지 전략들은 특수 도메인 적용, 다양한 데이터 처리, 규제 산업에서의 투명성 확보라는 과제들을 해결한다. 문 접근법은 AutoML을 단순한 하이퍼파라미터 최적화에서 포괄적인 도메인 맞춤형 솔루션으로 발전시켜, 전문가들이 심층적인 AI 지식 없이도 고성능 머신러닝 시스템을 개발할 수 있게 한다.

1. AutoGluon과 같은 오토ML 프레임워크가 금융/산업 현장의 예측 정확도를 높이기 위해 스택킹과 블렌딩 기반의 멀티 앙상블 전략을 실제 적용 중에 있다.

Wei Yin <sup>a, b</sup>, Berna Kirkulak-Uludag <sup>c</sup>, Dongmei Zhu <sup>a</sup>, Zixuan Zhou <sup>a</sup>

Show more

+ Add to Mendeley Share Cite

<https://doi.org/10.1016/j.asoc.2023.110302>

Highlights

- Economic development and the real-estate development variables are influential in predicting the credit default risk in China
- The Max-Relevance and Min-Redundancy (MRMR) method is used for feature selection to verify the selection result
- Stacking ensemble model yields high performance in precision and recall

Financial Fraud Detection Using Explainable AI and Stacking Ensemble Methods

Fahad Almalki  
s44580241@students.tu.edu.sa Computer Science  
Department, College of Computers and Information  
Technology, Taif University, Al-Hawiyah, Taif 21944, Makkah,  
Saudi Arabia

Mehedi Masud  
mmasud@tu.edu.sa Department of Computer Science, Taif  
University, Al-Hawiyah, Taif 21944, Makkah, Saudi Arabia

(May 2025)

Abstract

Losses from credit card fraud increased significantly worldwide from \$28.4 billion in 2020 to \$33.5 billion in 2022. However, traditional machine learning models often prioritize predictive accuracy, often at the expense of model transparency and interpretability. The lack of transparency makes it difficult for organizations to comply with regulatory requirements and gain stakeholders' trust. In this research, we propose a fraud detection framework that combines a stacking ensemble of well-known gradient boosting models: XGBoost, LightGBM, and CatBoost. In addition, explainable artificial intelligence (XAI) techniques are used to enhance the transparency and interpretability of the model's decisions. We used SHAP (SHapley Additive Explanations) for feature selection to identify the most important features. Further efforts were made to explain the model's predictions using Local Interpretable Model-Agnostic Explanation (LIME), Partial Dependence Plots (PDP), and Permutation Feature Importance (PFI). The IEEE-CIS Fraud Detection dataset, which includes more than 590,000 real transaction records, was used to evaluate the proposed model. The model achieved a high performance with an accuracy of 99% and an AUC-ROC score of 0.99, outperforming several recent related approaches. These results indicate that combining high prediction accuracy with transparent interpretability is possible and could lead to a more ethical and trustworthy solution in financial fraud detection.

- 1. 서동준, 정경용. (2025). AutoML 프레임워크에서 AI모델 성능 고도화를 위한 기술 전략. 한국정보기술학회논문지
- 2. Stacking ensemble method for personal credit risk assessment in Peer-to-Peer lending.
- 3. Financial Fraud Detection Using Explainable AI and Stacking Ensemble Methods

3. P2P 대출 플랫폼 등에서 스택킹 앙상블을 적용해 신용 위험(부도 가능성) 예측 정확도를 높이는 연구와 실증 사례