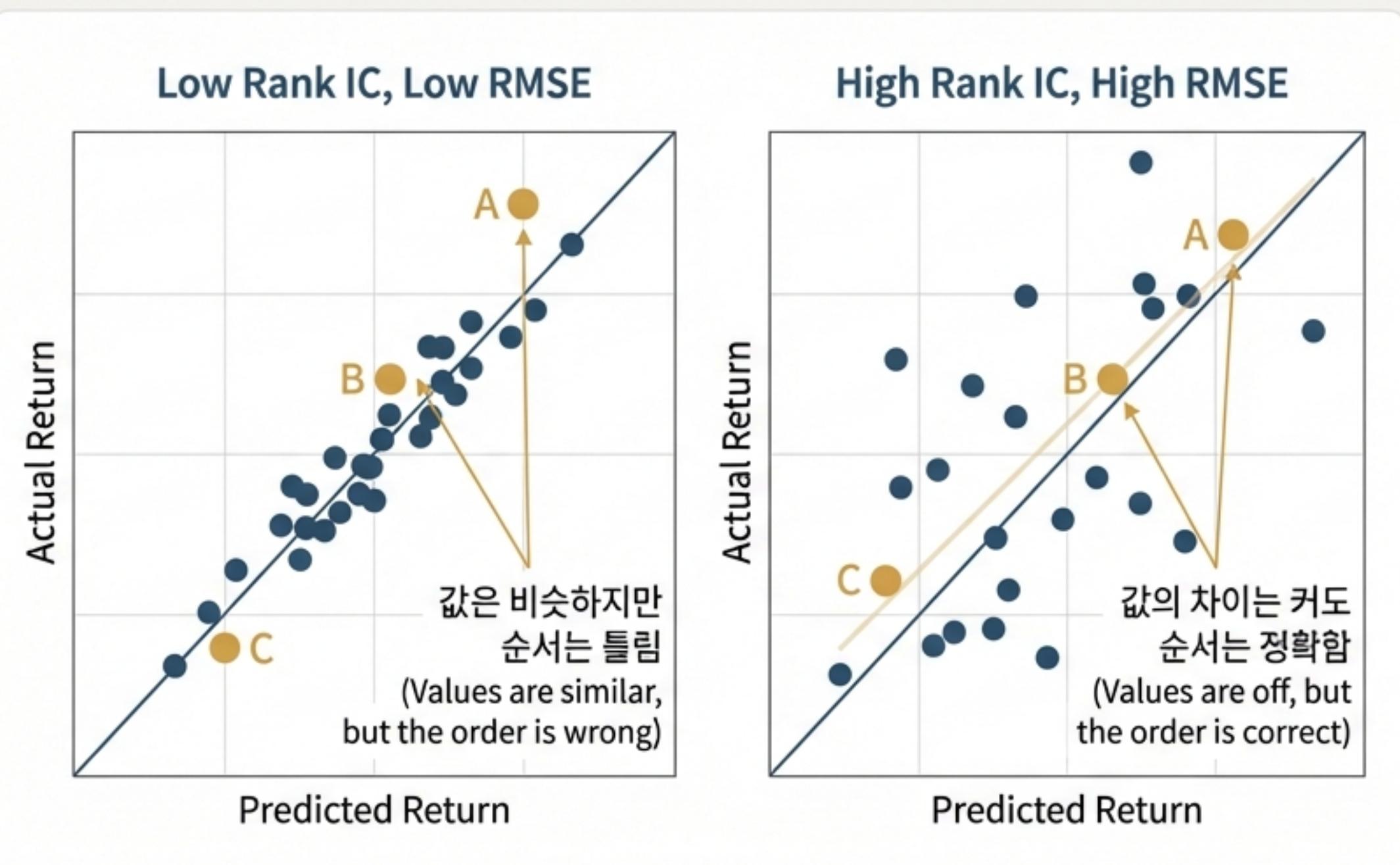


퀀트 랭킹 예측의 기술: Kaggle 주식 수익률 예측 대회 공략기

Baseline에서 Top Score까지, Rank IC 0.089 달성 전략 분석

The Mission: 무엇을, 그리고 '어떻게' 예측해야 하는가?

- **대회 목표:** 중국 A주(A-Shares)의 미래 수익률 상대 순위 예측.
 - **핵심 평가 지표:** Rank IC (Information Coefficient).
이는 예측값의 '순위'와 실제 수익률의 '순위' 간의 Spearman Correlation을 측정합니다.
 - **근본적인 차이:** 우리의 목표는 "주식 A가 5% 오를 것이다"라고 맞추는 것이 아니라, "주식 A가 주식 B보다 더 많이 오를 것이다"라는 순서를 맞추는 것입니다.



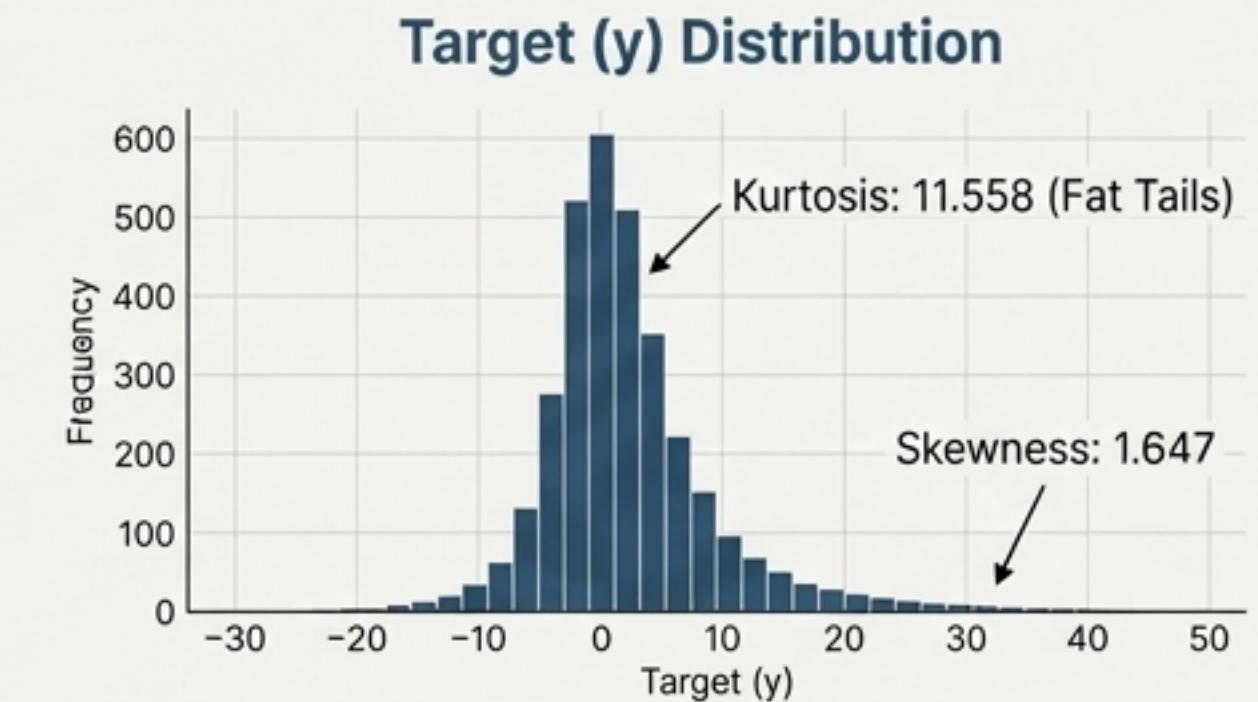
The Arena: 데이터의 특성과 마주한 4가지 난관

데이터 개요

- 학습 데이터: 470만 행, 익명화된 Alpha Factors ($f_0 \sim f_6$)
- 시간 범위: 학습 (date 0-1701), 테스트 (date 1702-2803)

주요 도전 과제

- 약한 신호 (Weak Signal):** 피처와 타겟(y)의 상관관계가 극히 낮음. 최대 상관계수: **0.0417** (f_5).
- Fat-tail 분포:** 타겟 변수(y)의 Skewness **1.647**, Kurtosis **11.558**. 극단적인 수익/손실이 많아 예측이 불안정함.
- Cold Start 문제:** 테스트 데이터에만 존재하는 **1,526**개의 신규 주식.
- Look-ahead Bias 위험:** 시계열 데이터에서 미래 정보를 참조하여 발생하는 과적합 가능성.



Stock Overlap (Train vs. Test)



시작점: 전통적 회귀 접근법 (Baseline)

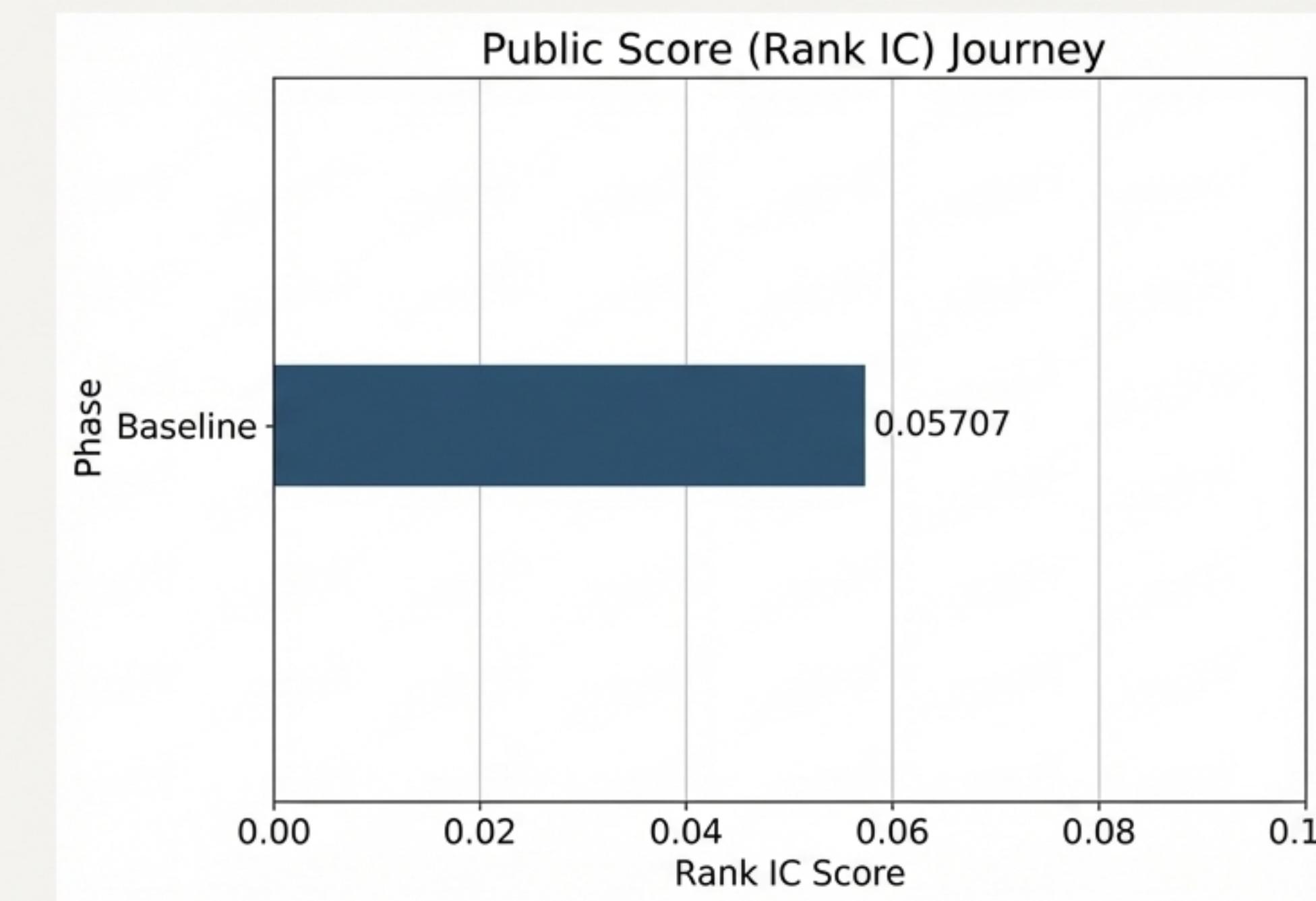
핵심 전략:

- 모델: LightGBM Regressor
- 목표 함수 (Objective): RMSE (평균 제곱근 오차). 모델이 실제 수익률 '값'을 가장 근접하게 예측하도록 학습.

피처 엔지니어링:

- 이동평균 ($f_{0_ma_5}$, $f_{0_ma_10}$)
- 종목/섹터별 통계량 ($stock_mean_f_0$, $sector_mean_f_0$)

```
lgbm_params = {  
    'objective': 'rmse',  
    ...  
}
```



점진적 개선: 평가 지표를 향한 첫걸음

시도 1: Rank-aware 피처 추가

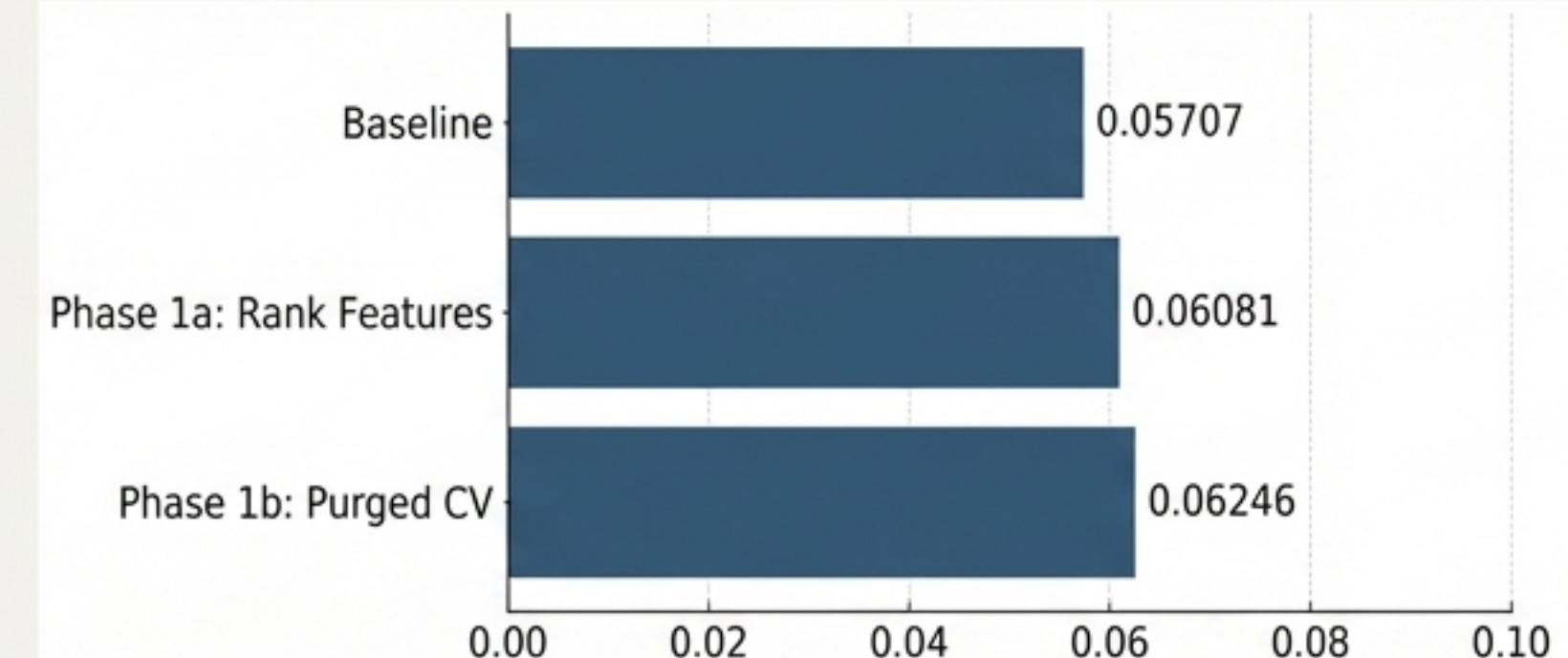
- 전략: 각 날짜 내에서 피처의 상대적 순위를 계산한 'Cross-Sectional Rank' 피처 (e.g., rank_f_4) 추가. 평가 지표인 Rank IC에 직접적인 힌트를 제공.
- 결과: Score **0.06081** (+6.6%)

시도 2: 검증 전략 강화

- 전략: Purged Group Time Series Split 도입. Train/Validation set 사이에 20일의 공백(gap)을 두어 미래 정보 누수(Look-ahead Bias)를 원천 차단.
- 결과: Score **0.06246** (+9.4%)

교훈: "피처와 검증 방식의 개선은 유효했으나, 성능의 '점프'를 이끌어내지는 못했다."

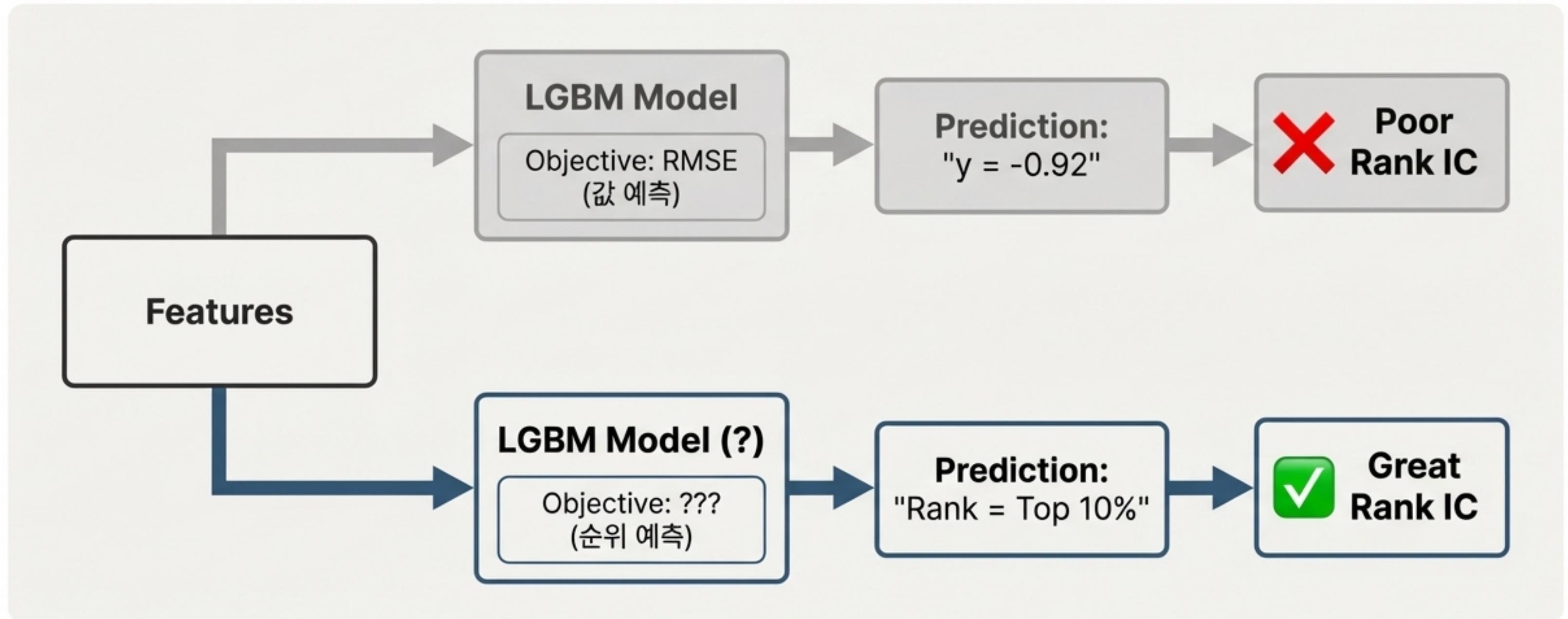
Public Score (Rank IC) Journey



Purged Cross-Validation



우리는 잘못된 문제를 풀고 있었다: 목표와 수단의 불일치



모델은 값의 정확성을 위해 학습하고 있었지만, 평가는 순서의 정확성으로 이루어졌다.
이 근본적인 불일치(Misalignment)가 성능의 발목을 잡고 있었다.

The Breakthrough: 문제 자체를 변환하다 (Target Transformation)

핵심 아이디어: 수익률 값(`y`)을 직접 예측하는 대신,
문제 자체를 '순위 예측 문제'로 재정의한다.

구체적인 방법:

- 각 날짜(`date`) 그룹 내에서
- 실제 수익률 `y` 값의 백분위 순위(Percentile Rank)를 계산
- 이 순위 값 (0.0 ~ 1.0)을 새로운 타겟 변수로 사용

Original Target → **Transformed Target**

Percentile Rank by Date

date	code	y (original)	
0	s_4394	-0.925	
0	s_4451	-0.568	
0	s_594	-0.747	

date	code	y (original)	y_rank_pct (new target)
0	s_4394	-0.925	0.231
0	s_4451	-0.568	0.452
0	s_594	-0.747	0.315

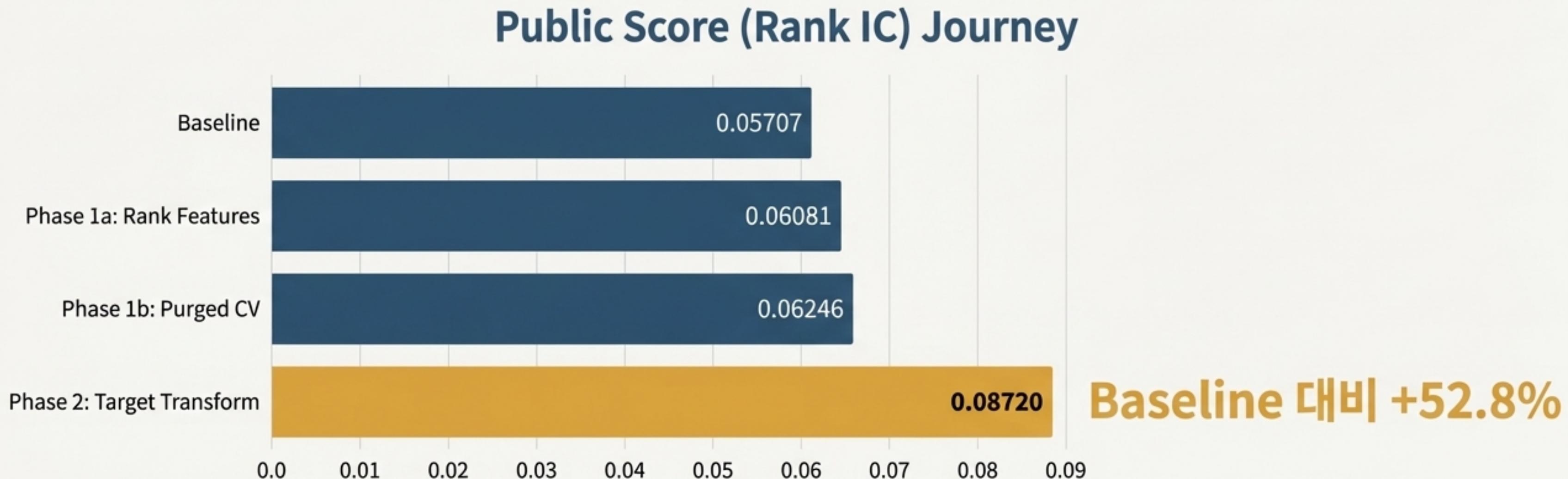
작동 원리:

이 변환을 통해 모델의 손실 함수(MSE on Ranks)가
평가 지표(Rank IC)와 직접적으로 **정렬(Align)**됩니다.
모델은 이제 '값'이 아닌 '순위'를 맞추는 방향으로 최적화됩니다.

Phase 2: 성능의 폭발적인 도약

적용 전략:

1. 핵심 동인: Target Rank Transformation
2. 안정성 강화: 5개의 다른 Random Seed로 학습 후 결과를 평균하는 5-Seed Ensemble 적용



Phase 3: 디테일을 통한 추가 성능 향상

전략 1: 상호작용 피처 (Interaction Features)

- 개념: 단일 변수가 아닌, 경제적 의미를 가질 수 있는 변수 간의 조합을 추가.
- 예시:

$$inter_price_tech = f_0 \times f_5$$

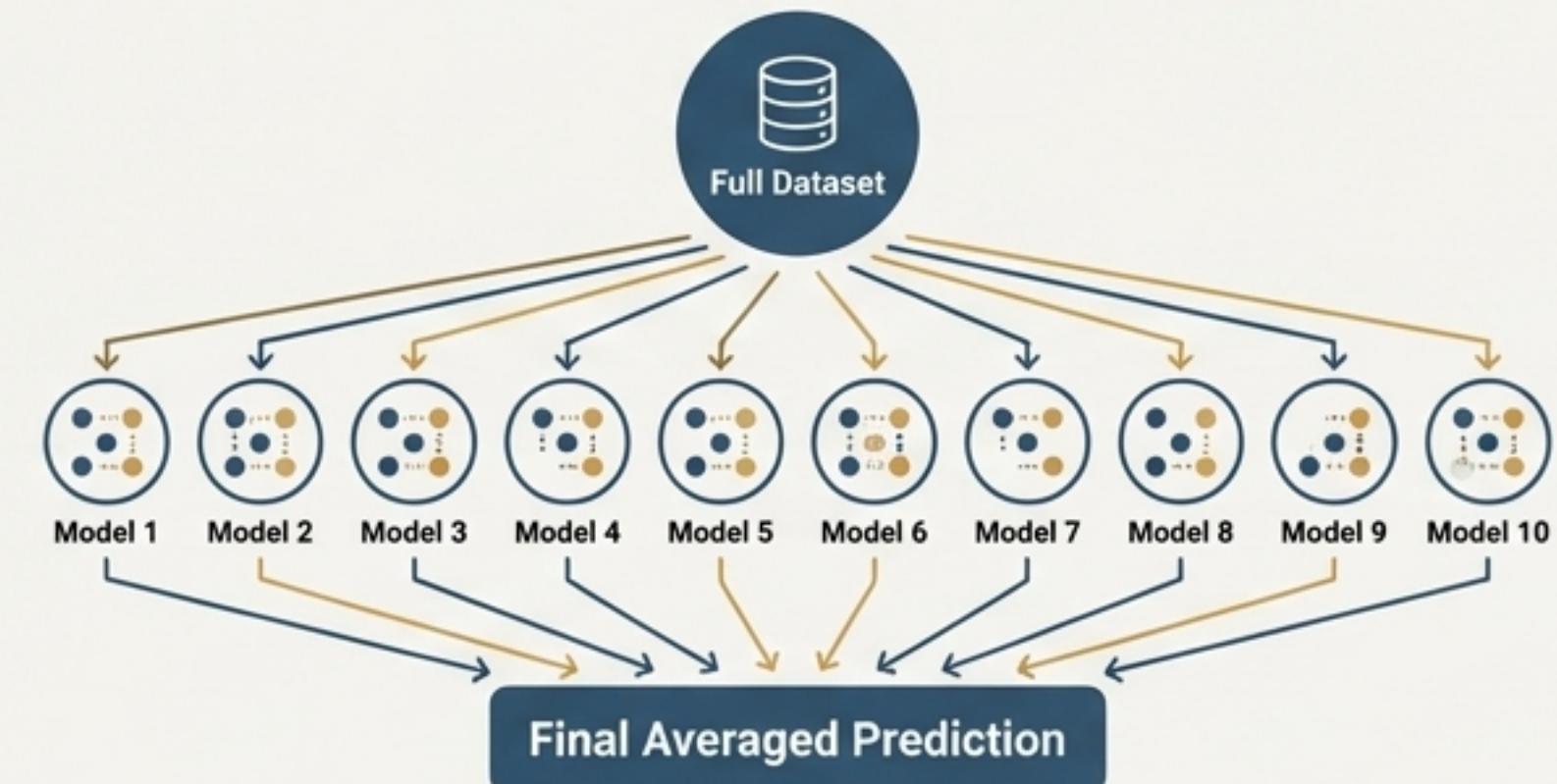
가격 관련 지표 × 기술 지표

$$inter_price_vol = f_0 \div f_4$$

가격 대비 거래량

전략 2: 양상을 강화 (Enhanced Ensemble)

- 개념: 모델의 분산을 더욱 줄이고 예측 안정성을 극대화.
- 방법: 양상을 규모를 5-Seed에서 **10-Seed Bagging**으로 확대.



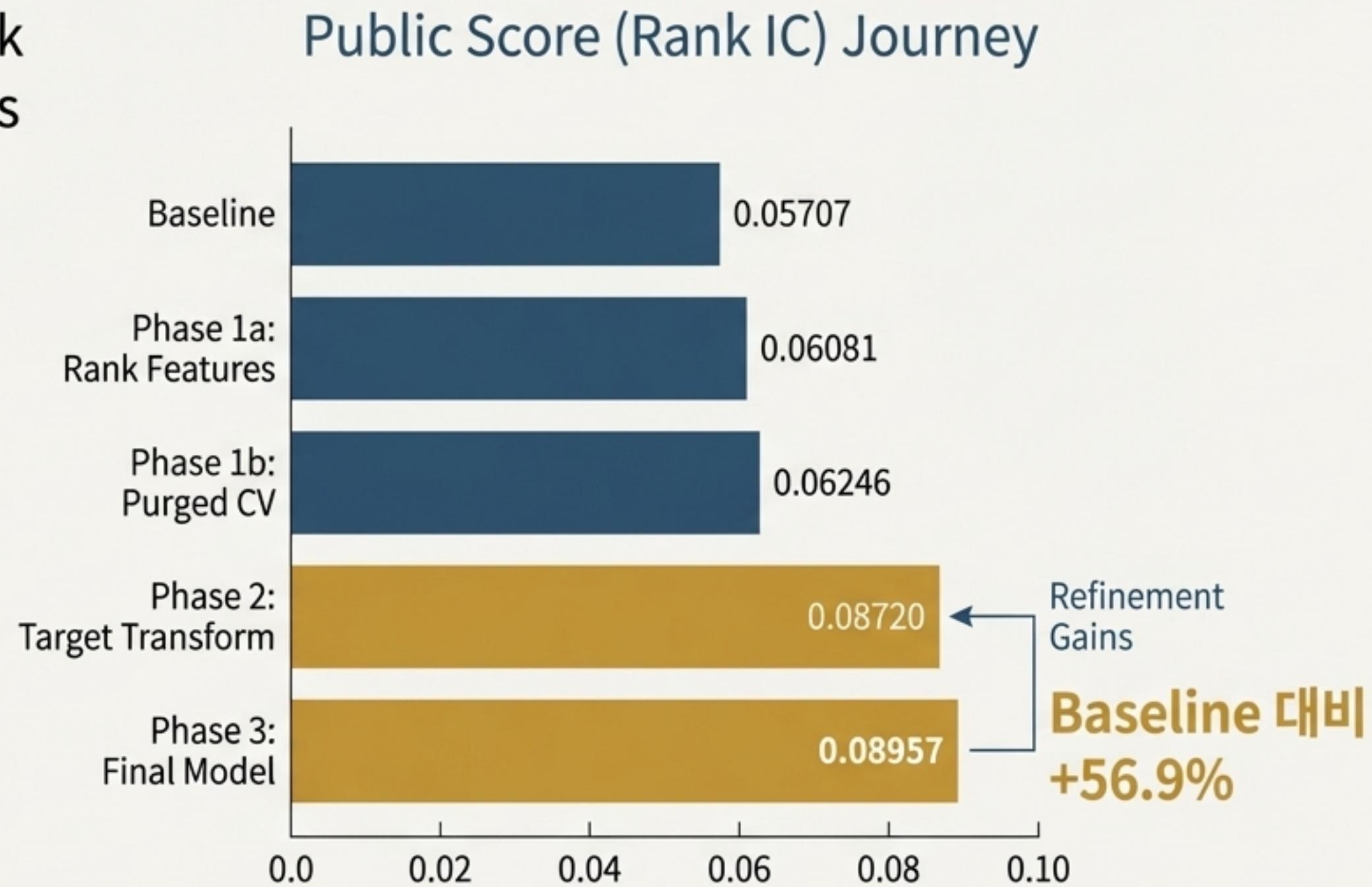
최종 성과: 순수 머신러닝 접근법의 정점

최종 모델 구성: LightGBM + Target Rank Transformation + Interaction Features + 10-Seed Bagging

최종 결과:

- * Public Score (Rank IC): **0.08957**
- * 총 성능 향상: Baseline 대비 **+56.9%**

통찰: 핵심 돌파구(Target Transform)
이후에는, 정교한 피처와 강력한 앙상블이
점수를 ‘쥐어짜는(Squeezing)’ 데
유효하다.



The Journey at a Glance: 모델 진화의 여정

Phase	Key Strategy	Public Score (Rank IC)	Improvement (vs Baseline)
Baseline	RMSE Objective	0.05707	-
Phase 1	Rank Features + Purged CV	0.06246	+9.4%
Phase 2	Target Rank Transform + 5 Seeds	0.08720	+52.8% (Key Driver)
Phase 3	Interaction Features + 10 Seeds	0.08957	+56.9% (Best)

핵심 성공 요인: 이 프로젝트에서 얻은 4가지 교훈



1. 목표와 수단을 정렬하라 (Align Your Tools with Your Goal)

손실 함수(Loss Function)와 평가 지표(Evaluation Metric)를 일치시키는 것이 다른 어떤 튜닝보다 중요하다.



2. 풀리지 않으면, 문제를 변환하라 (Transform the Problem)

직접적인 예측이 어려울 때, 타겟 변수를 변환(순위, 로그, 차분)하는 것은 가장 강력한 무기가 될 수 있다.



3. 검증 전략을 신뢰하라 (Trust Your Validation)

‘Purged CV’와 같은 엄격한 검증 체계는 리더보드 점수와의 괴리를 줄이고 모델의 일반화 성능을 보장한다.



4. 앙상블은 거의 항상 정답이다 (Ensemble is (Almost) Always the Answer)

다양한 시드(Seed)를 사용한 앙상블은 안정적인 성능 향상을 위한 가장 확실한 방법 중 하나다.

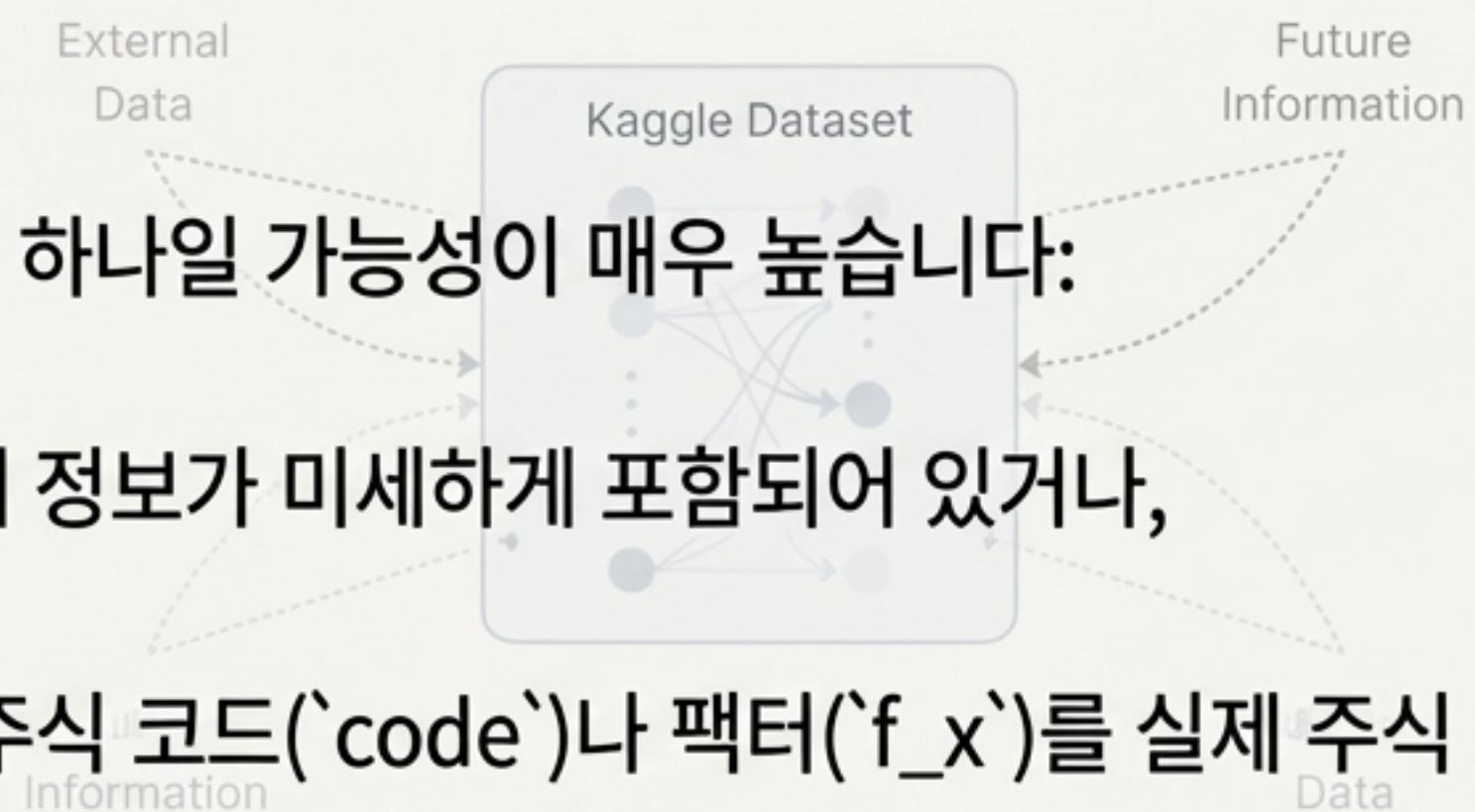
리더보드에 대한 고찰: 1.0에 가까운 점수는 무엇을 의미하는가?

관찰: 리더보드 상위권의 Rank IC 점수는 1.0에 근접합니다. 이는 금융 시계열 데이터에서 거의 불가능한 수준의 예측력입니다.

현실적인 추론: 이러한 수치는 다음 두 가지 시나리오 중 하나일 가능성이 매우 높습니다:

- 정보 누수 (Data Leakage):** 훈련 데이터셋에 미래 정보가 미세하게 포함되어 있거나, 데이터셋 자체의 결함을 이용.
- 비식별화 해제 (De-anonymization):** 익명화된 주식 코드(`code`)나 팩터(`f_x`)를 실제 주식 및 데이터와 매칭하여 외부 정보를 활용.

본 프로젝트의 의의: 우리는 데이터 유출이나 외부 정보 없이, 주어진 데이터 내에서 순수 머신러닝 기법만으로 달성 가능한 현실적인 성능의 상한선을 탐색했습니다.



향후 연구 방향 및 핵심 참고 자료

Next Steps for Improvement



Advanced Features: Genetic Programming (gplearn 라이브러리 등)을 활용한 수식 기반 피처 자동 생성.



Model Diversity: AutoML (e.g., TabNet, AutoGluon)을 활용하여 GBDT 외의 다양한 모델을 앙상블에 추가.

Key Concepts & Resources



Purged Cross-Validation: M. López de Prado



Information Coefficient (IC): The Fundamental Law of Active Management



Alpha Factors: WorldQuant, "101 Formulaic Alphas"



Learning to Rank (LTR): XGBoost/LightGBM의 `lambdarank` objective

가장 큰 도약은 복잡한 모델이 아닌, 문제에 대한 깊은 이해에서 온다.

- 이 프로젝트의 여정은 '더 나은 모델'을 찾는 과정이 아니라, '**올바른 문제**'를 정의하는 과정이었습니다.
- Baseline(0.057)에서 Phase 1(0.062)까지의 튜닝보다, 단 하나의 아이디어(Target Transformation)가 Phase 2(
- 성공적인 머신러닝 프로젝트의 핵심은 알고리즘이 아닌, '**손실 함수, 평가 지표, 그리고 비즈니스 목표**' 이 세 가지를 일치시키는 능력에 있습니다. 알고리즘이 아닌, '**손실 함수, 평가 지표, 그리고 비즈니스 목표**' 폭발 (0.087)

