SECOM 불량 예측 모델링 보고서

1. 프로젝트 개요

- SECOM 데이터(590개 센서, 1,567개 웨이퍼) 기반으로 Pass/Fail 불량을 조기 감지하는 분류 모델을 구축하였다.

- 사용자는 ROC-AUC 0.75 이상을 목표로 하였으며, SMOTE 기반 불균형 처리와 하이퍼파라미터 탐색을 포함한 전체 실험 자동화를 요구하였다.

2. 데이터 및 전처리

- 입력 데이터: `uci-secom.csv`, 레이블: `secom.names`/`secom_labels.data`.

- 결측치는 중앙값(SimpleImputer)으로 보간하고, 분산이 0에 근접한 특성은 VarianceThreshold로 제거하였다.

- Stratified 80/20 Train-Test 분할 후, 모든 모델은 imblearn `Pipeline`을 통해 SMOTE→StandardScaler→Estimator 순으로 학습하여 데이터 누수를 방지하였다.

3. 실험 방법

- XGBoost: Particle Swarm Optimization(입자 10, 반복 10)으로 학습률, 최대 깊이, 추정기 수, 정규화 계수 등을 탐색하였다. 최적 해는 learning_rate 약 0.087, max_depth=10, n_estimators 약 600, reg_lambda 약 3.8 등으로 수렴하였다. Seed를 5개(42~82)로 달리한 앙상블을 구성하고 확률 평균을 사용하였다.

- LightGBM: 120회 RandomizedSearchCV(3-Fold) 후 상위 3개 파라미터 셋을 다시 3개 시드로 적합하여 9모델 앙상블을 구성했다. 모든 모델에서 class_weight를 scale_pos_weight에 맞추어 적용했다.

- Threshold Tuning: 두 앙상블 및 이후 혼합 모델의 예측 확률에 대해 0.05~0.95(0.05 step) 범위를 탐색하며 Fail 클래스 F1이 최대가 되는 임계값을 선택하였다.

- 혼합 시도: XGBoost 확률과 LightGBM 확률을 단순 가중 평균(가중치 0.0~1.0, 0.05 step)으로 섞어 ROC-AUC를 비교하였다.

- 추가 시도: HistGradientBoosting RandomizedSearch(120회, 5-Fold)를 실행했으나 연산 시간이 매우 길어 사용자 중단(KeyboardInterrupt)으로 종료되었다.

4. 테스트 세트 성능 비교

- XGBoost 앙상블: ROC-AUC 0.7179. 기본 임계값 0.5에서는 Fail Recall 0.0이었으나, 최적 임계값 0.10 적용 시 Accuracy 0.8949, Fail F1 0.2667, Fail Recall 0.2857, Fail Precision 0.2500.

- LightGBM 앙상블: ROC-AUC 0.7377, 최적 임계값 0.10에서 Accuracy 0.8694, Fail F1 0.3692, Fail Recall 0.5714, Fail Precision 0.2727. ROC-AUC는 약간 향상, 재현율은 크게 개선되었다.

- XGBoost-LightGBM 블렌딩: XGB 비중 0.00(사실상 LightGBM 단독)에서 ROC-AUC 0.7463으로 최고점을 기록했으나, 현재 노트북 로직의 `current_best_auc` 갱신 조건 미충족으로 최종 모델로 고정되지는 않았다. 추가 임계값·리포트 갱신 필요.

5. 분석 및 한계

- ROC-AUC 0.7463으로 목표(0.75)에 근접했지만 아직 미충족이다. 보다 세밀한 가중 탐색 또는 스태킹 메타모델이 유효할 수 있다.

- Fail 클래스 표본이 21개에 불과해 재현율 변동성이 매우 높다. K-fold 내 클래스 불균형이 심하여 CV와 테스트 간 편차가 발생한다.

- HistGradientBoosting 탐색이 중단되어 비교군이 제한적이며, LightGBM도 9개 모델 앙상블에 그쳐 탐색 공간이 상대적으로 좁다.

6. 향후 권장 사항

- 혼합 확률에 대해 더 촘촘한 가중 격자(예: 0.00~1.00, 0.01 step) 또는 베이지안 최적화를 적용하고, Fail-F1 기준 임계값·혼동행렬을 다시 산출한다.

- HistGradientBoosting 탐색 시 `HalvingRandomSearchCV`나 `n_iter` 축소를 사용해 중단 없이 완료하고, CatBoost 등 다른 그라디언트 부스팅 계열도 검토한다.

- SMOTE 파라미터(k_neighbors 등)와 Threshold 탐색 범위를 동시에 최적화하여 재현율과 정밀도의 균형을 맞춘다.

- 최종 리포트를 위해 ROC 곡선, PR 곡선, 특성 중요도 시각화를 함께 저장하고, 생산 환경에서는 안정적인 seed 고정 및 모델 모니터링 절차를 포함한다.

7. 시각 자료

- `reports/performance_summary.png`: 주요 모델(XGBoost 앙상블, LightGBM 앙상블, XGB/LGB 블렌드)의 테스트 세트 ROC-AUC·정확도·Fail 클래스 지표를 표 형태로 요약한 이미지이며, 본 PDF에 첨부된다.

SECOM 모델 테스트 성능 요약 (Test Set)

모델	ROC-AUC	Accuracy	Fail F1	Fail Recall	Fail Precision
XGBoost 앙상블	0.7179	0.8949	0.2667	0.2857	0.25
LightGBM 앙상블	0.7377	0.8694	0.3692	0.5714	0.2727
XGB/LGB 블렌드	0.7463	-	-	-	-

* 블렌드 모델은 ROC-AUC만 산출 (Fail-F1 기반 임계값 재적용 필요)