ANALYSIS REPORT

당뇨병 예측 모델링: 통계분석 및 머신러닝 접근

Kaggle Playground Series S5E12 (Diabetes Prediction Challenge)

- **데이터** 합성 당뇨 예측 데이터 (Train/Test Dataset)
- **목표** 진단 확률 예측 및 위험 요인에 대한 통계적 가설 검정
- **작성일** 2025년 12월 26일
- **출처** Kaggle S5E12, ML분석.pdf, 통계분석.html

Table of Contents

본 발표의 주요 구성 및 분석 흐름

01

통계분석

Statistical Analysis

- 타깃 변수 및 분석 방법론 개요
- HbA1c(당화혈색소) 다중회귀 분석
- 당뇨 진행 단계 순서형 로지스틱
- 인슐린 수치 ANOVA 검정 결과

SECTION 1 — 5 Slides

02

머신러닝 분석

Machine Learning Modeling

- 모델 선정 및 평가 전략 수립
- 알고리즘별 성능 비교 평가
- 최우수 모델 (LightGBM) 심층 분석
- Kaggle 리더보드 제출 및 점수

SECTION 2 — 5 Slides

03

인사이트 및 제언

Insights & Conclusion

- 위험 요인 및 보호 요인 종합
- 모델 성능 향상을 위한 제언
- 향후 연구 및 개선 방향

SECTION 3 — 1 Slide

SECTION 01 STATISTICAL ANALYSIS

통계분석

가설 검정을 기반으로 당뇨병 발병 및 진행과 관련된
핵심 지표들의 통계적 유의성을 도출하고 해석합니다.

HBA1C
다중 선형 회귀 분석

DIABETES STAGE
순서형 로지스틱 회귀

INSULIN LEVEL
일원 분산 분석 (AN

01

통계분석 개요 및 데이터 구조

분석 대상 및 방법론 (Targets & Methodology)

종속변수 (DEPENDENT VARIABLE)	DATA TYPE	통계 분석 방법 (STATISTICAL METHOD)
HbA1c 당화혈색소 수치	연속형	**다중 선형 회귀 분석** Multiple Linear Regression
diabetes_stage 정상 → 전당뇨 → 당뇨	순서형	**순서형 로지스틱 회귀** Ordered Logit Regression
insulin_level 인슐린 수치 (Log 변환)	연속형	**일원 분산 분석 (ANOVA)** Analysis of Variance

주요 공변량 (Key Covariates)

모델에 투입된 주요 예측 변수:

- 당뇨 가족력
- 연령
- 신체활동
- 식단 점수
- BMI
- 허리-엉덩이비(WHR)
- 중성지방
- 소득 수준
- 교육 수준

분석 목적 (Objective)

"임상 데이터셋 내에서 혈당(HbA1c), 질병 진행 단계, 인슐린 분비능에 유의미한 영향을 미치는 위험 인자 및 보호 요인을 식별하고, 그 통계적 유의성을 검증한다."

HbA1c: 다중 선형 회귀 분석 결과

Target Variable
HbA1c (Glycated Hemoglobin)

주요 통계적 유의성 (Statistical Significance)

위험 인자 (Risk Factors) `p < 0.001`

당뇨 가족력과 연령이 혈당 상승의 가장 강력한 예측 인자로 확인됨.

- 가족력 (Coef: +0.475): 유전적 요인의 지배적 영향
- BMI 및 중성지방: 대사 지표 악화 시 HbA1c 동반 상승

보호 요인 (Protective Factors) `Negative Coef`

신체 활동과 건강한 식단은 혈당 수치를 낮추는 유의미한 효과를 보임.

- 생활습관 개선(운동, 식단)이 혈당 관리의 핵심 통제 변수임

비유의 요인 (Non-Significant) `p > 0.05`

소득 수준(Income), 인종(Race), 교육 수준(Education) 등 사회경제적 변수는 HbA1c 수치와 통계적으로 유의한 선형 관계가 관찰되지 않음.

INTERPRETATION

유전적 소인(가족력)은 통제 불가능하나, 신체 활동량 증대와 식단 관리는 HbA1c를 낮추는 가장 효과적인 중재 방안입니다.

요인별 상대적 영향력 (Standardized Coefficients)

Method: OLS Regression

당뇨 진행 단계: 순서형 로지스틱 회귀 분석

Target Variable
Diabetes Stage (Ordered)

주요 오즈비 분석 (Odds Ratio Analysis)

핵심 위험 인자 (Critical Risk Factors) — OR > 1.0

당뇨 가족력은 질병 단계 진행의 가장 절대적인 위험 요인이며, 단순 비만(BMI)보다 복부 비만(WHR)이 더 치명적임.

- 가족력 (OR 3.21): 상위 단계 진행 확률 3배 이상 증가
- 허리-엉덩이 비율(WHR) (OR 1.68): 내장 지방의 위험성 시사

보호 요인 (Protective Factors) — OR < 1.0

건강한 식단과 규칙적인 신체 활동은 당뇨병 진행을 억제하는 보호 효과가 확인됨.

- 식단 점수 (OR 0.96): 식습관 개선 시 위험도 약 4% 감소
- 신체 활동 (OR 0.997): 미세하지만 통계적으로 유의한 억제 효과

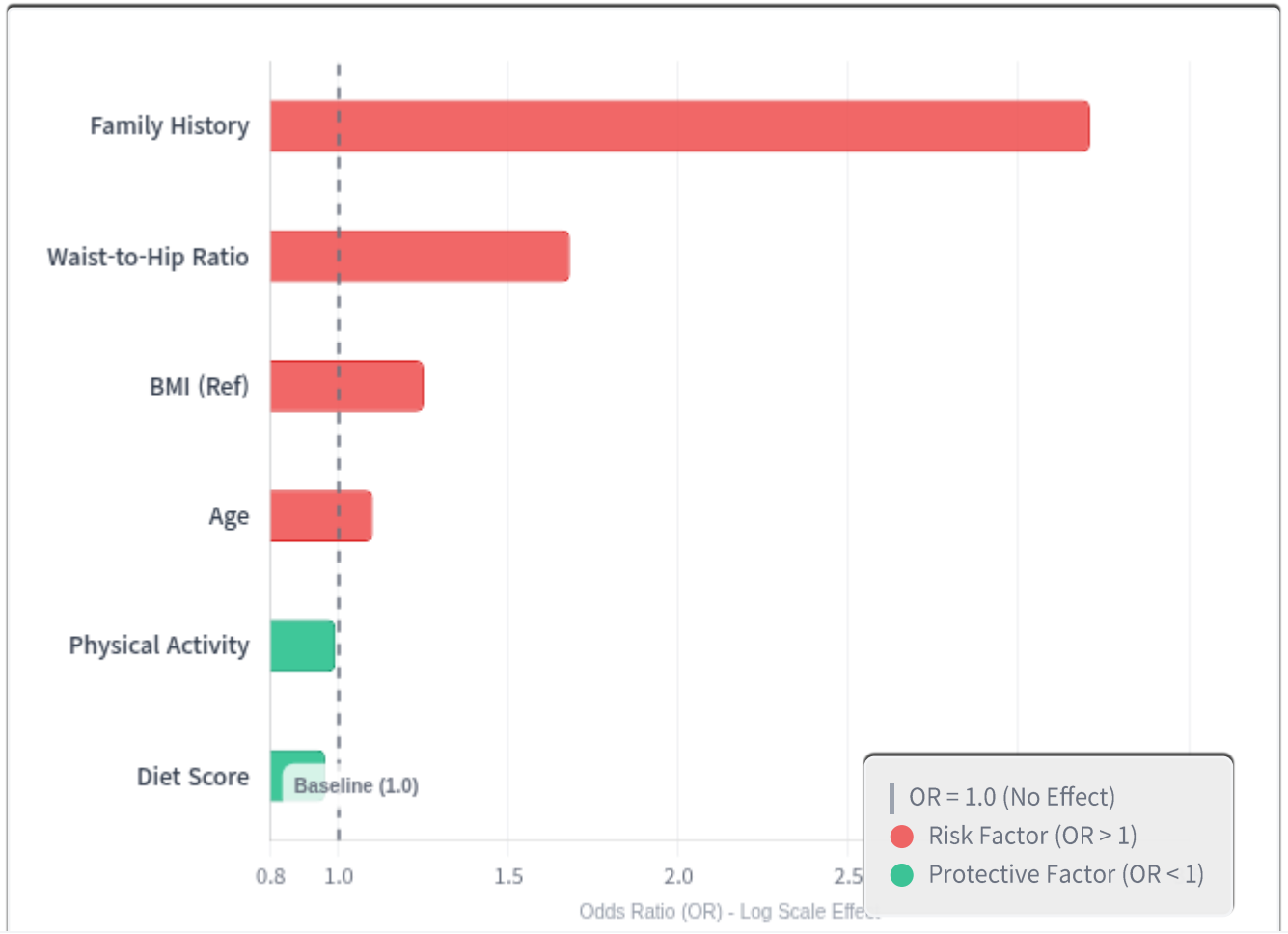
해석 (Interpretation)

유전적 요인이 기저 위험을 결정하지만, 복부 비만 관리(WHR 감소)가 질병 진행을 막는 가장 효과적인 수정 가능 목표임을 시사함.

KEY INSIGHT

주요 변수별 오즈비 (Odds Ratio)

Model: Ordered Logit



인슐린 수치: 일원 분산 분석(ANOVA) 결과

Target Variable
Insulin Level (Log-transformed)

통계적 비유의성 (Non-Significance)

사회경제적 요인 (Socio-economic) — Not Significant

소득 수준(p=0.19)과 교육 수준(p=0.95)은 인슐린 수치 차이에 통계적으로 유의미한 영향을 주지 않음.

- 소득이 높다고 해서 인슐린 저항성이 낮다는 증거 없음
- 교육 수준에 따른 식습관 차이가 인슐린 수치로 직결되지 않음

인구통계학적 요인 (Demographic) — Not Significant

성별(p=0.34)과 인종(p=0.16)에 따른 인슐린 수치의 유의미한 평균 차이는 관찰되지 않음.

- 생물학적 성별 차이나 인종적 특성이 직접적 원인이 아님

기타 요인 (Others) — Not Significant

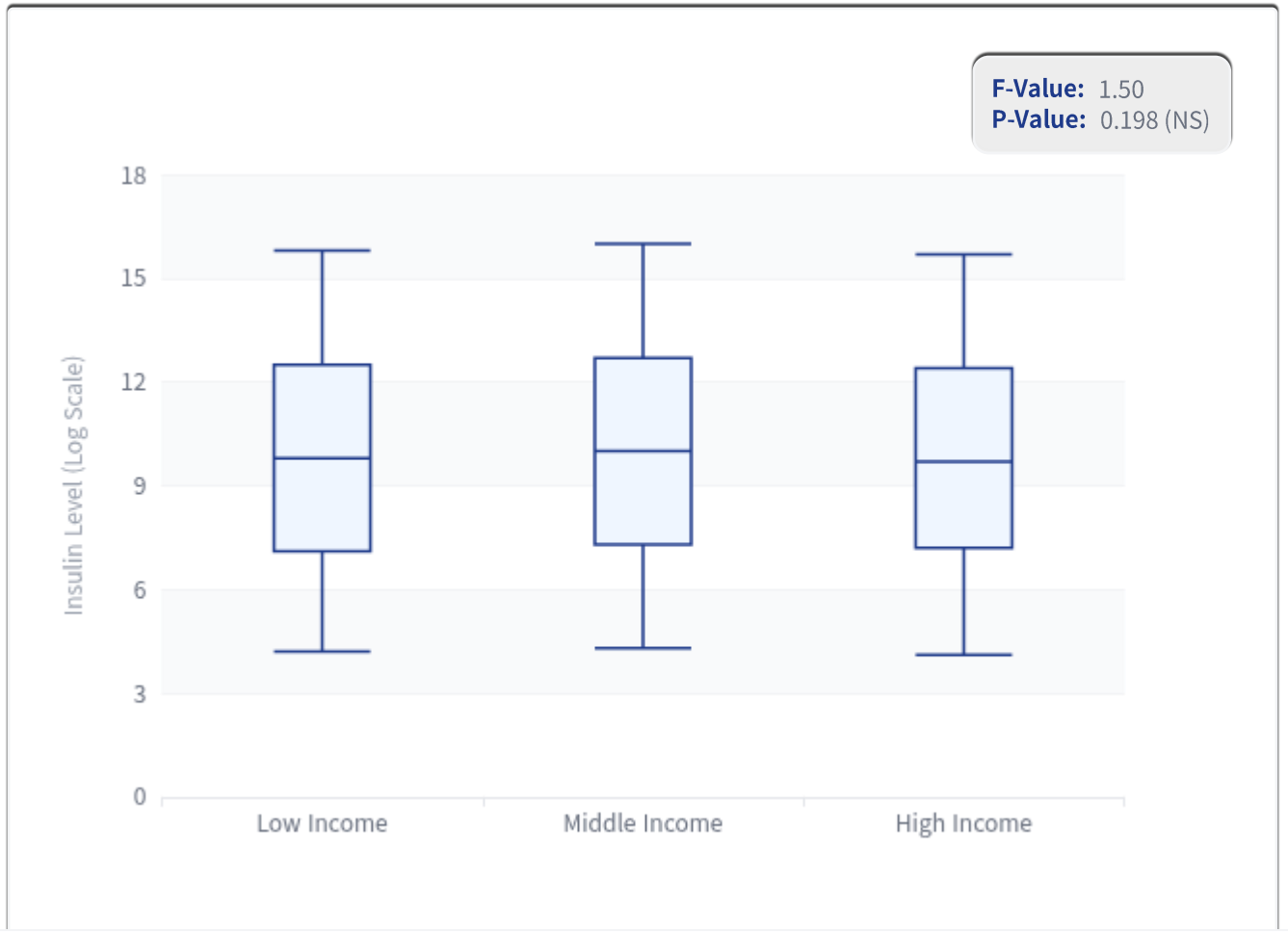
흡연 상태(p=0.53) 또한 인슐린 수치와 뚜렷한 연관성이 발견되지 않았음.

ANALYTIC INSIGHT

인슐린 수치는 개인의 사회경제적 배경이나 단순 인구통계학적 특성보다는, 비만도(BMI, WHR)나 유전적 요인 등 생물학적 지표에 더 밀접하게 반응하

소득 수준별 인슐린 수치 분포 (Distribution)

Test: One-way ANOVA



SECTION 02 MACHINE LEARNING ANALYSIS

머신러닝 분석

당뇨병 진단 여부를 정확히 예측하기 위해
다양한 머신러닝 모델을 구축하고 최적의 성능을 도출합니다.

MODELING STRATEGY
4가지 모델 및 RandomizedSearchCV

PERFORMANCE
ROC-AUC 중심 성능 비교

BEST MODEL
LightGBM 7252)

02

모델링 전략 및 실험 설계

예측 모델 및 전처리 (Models & Preprocessing)

CATEGORY	DETAILS & LIBRARIES
Baseline Models 기본 성능 확인	Linear Logistic Regression Tree Decision Tree Classifier
Ensemble Models 고성능 부스팅 알고리즘	Boosting XGBoost (Extreme Gradient Boosting) Boosting LightGBM (Light Gradient Boosting)
Preprocessing Column Transformer	수치형: StandardScaler (표준화) 범주형: OneHotEncoder / OrdinalEncoder

검증 전략 (Validation)

Stratified K-Fold CV

타겟 클래스 비율을 유지하며 5-Fold 교차 검증 수행 (n_splits=5). 데이터 불균형을 고려한 안정적인 평가.

PRIMARY METRIC

ROC-AUC Score

보조 지표: Accuracy, Precision, Recall, F1

최적화 (Optimization)

RandomizedSearchCV

광범위한 파라미터 공간을 효율적으로 탐색.

모델 성능 비교 평가 (Validation)

Validation Metric
ROC-AUC Score

MODEL NAME	ROC-AUC ★	ACCURACY	PRECISION	RECALL	F1 SCORE
Logistic Regression Baseline Linear	0.6940	0.5952	0.7540	**0.7591**	0.6653
Decision Tree Basic Tree	0.6917	0.6267	0.6205	0.6902	0.6531
XGBoost Gradient Boosting	**0.7241**	**0.6542**	**0.7734**	0.6285	0.6934
LightGBM Best Model	**0.7252**	**0.6537**	**0.7720**	**0.6318**	**0.6949**

최종 선정 모델: LIGHTGBM

4개 모델 중 ROC-AUC (0.7252) 점수가 가장 높으며, F1 Score(0.6949)에서도 우수한 성능을 보여 최종 모델로 선정하였습니다. XGBoost와 성능 차이는 미미했으나, 학습 속도와 일반화 성능 측면을 고려하였습니다.

성능 분석 인사이트

선형 모델(Logistic Regression)은 재현율(Recall: 0.7591)이 가장 높았으나, 전반적인 예측 정확도와 구분력(AUC)에서는 트리 기반 앙상블 모델(XGB, LGBM)이 비선형 패턴을 더 효과적으로 학습했습니다.

최우수 모델 선정: LightGBM

Validation Metric
ROC-AUC: 0.7252

모델 선정 근거 및 최적화

최고 예측 성능 (Best Performance)

Rank 1

LightGBM은 검증 데이터셋에서 ROC-AUC 0.7252를 기록하여 Logistic Regression(0.694) 및 Decision Tree(0.692) 대비 우수한 성능을 입증함.

XGBoost(0.7241)와 근소한 차이이나 계산 효율성 측면에서 우위

최적화 전략 (Optimization)

RandomizedSearchCV

교차 검증(Stratified K-Fold)을 통한 하이퍼파라미터 튜닝 수행.

Params: learning_rate, n_estimators, num_leaves
Objective: 과적합 방지(subsample) 및 일반화 성능 극대화

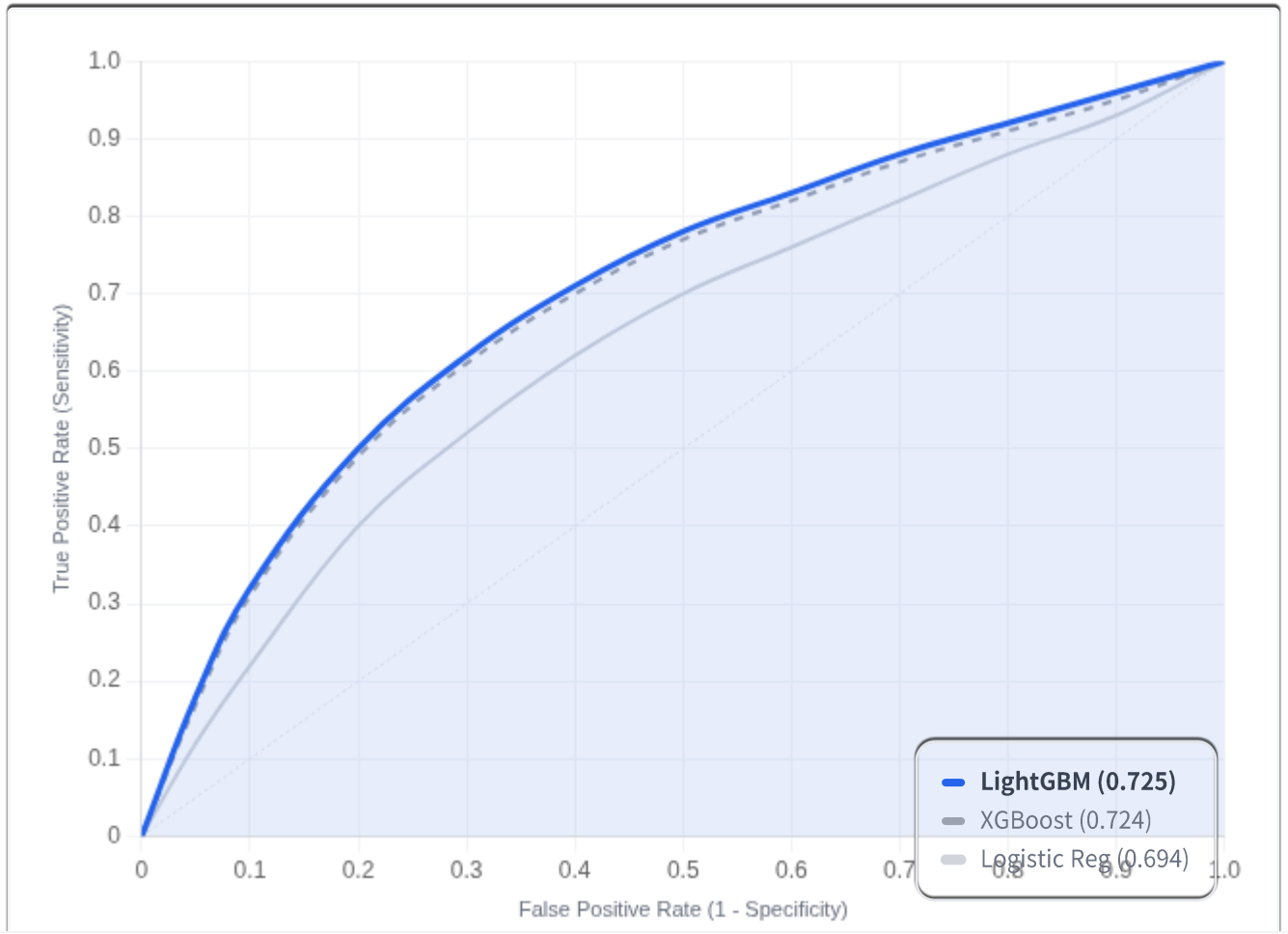
선정 이유 (Why LightGBM?)

Leaf-wise 트리 성장 방식을 사용하여 복잡한 비선형 패턴과 변수 간 상호작용을 효과적으로 포착하며, 대규모 데이터셋에서도 빠른 학습 속도를 보장함.

CONCLUSION

ROC Curve 비교 (Validation Set)

Sensitivity vs (1 - Specificity)



캐글 제출 및 최종 점수 (Submission & Final Score)

제출 프로세스 (Submission Pipeline)

LightGBM Model
최적화된 모델 로드

Predict Probability
예측 확률 산출 (Not Label)

Submission.csv
ID, diagnosed_diabetes

평가 지표 (Evaluation Metric)

ROC-AUC (Area Under the Curve)

이 대회는 이진 분류(Binary Classification) 문제로, 모델이 양성 클래스(당뇨병 발병)를 얼마나 잘 구분하는지를 평가합니다. 단순 정확도(Accuracy)보다 불균형 데이터 및 확률 예측 성능 평가에 적합합니다.

리더보드 점수 (Scores)

INTERNAL VALIDATION

0.7252

5-Fold Stratified CV Average

Public LB	Pending
Private LB	Hidden

검증 점수와 유사한 수준 예상

향후 액션 (Next Steps)

1 전체 Train 데이터로 최종 모델 재학습

2 Test 데이터에 대한 확률 예측 생성

3 Kaggle 리더보드 제출 및 점수 기록

핵심 분석 결과 (Key Insights)

위험 및 보호 요인 식별

통계적 유의성 검증 완료

핵심 위험 인자 (Risk Factors)

가족력, 연령, 복부비만(WHR)이 당뇨 진행의 가장 강력한 예측 변수임이 확인됨.

주요 보호 요인 (Protective Factors)

규칙적인 신체 활동과 식단 관리는 질병 진행을 억제하는 유의미한 효과를 가짐.

최적 모델 선정

LightGBM 모델 우수성 확인

선형 모델(Logistic Regression) 대비 LightGBM이 변수 간의 복잡한 비선형 관계와 상호작용을 더 효과적으로 포착하여 가장 우수한 성능을 기록함.

BEST PERFORMANCE

Validation ROC-AUC **0.7252**

향후 개선 로드맵 (Future Improvements)

특성 공학 고도화

단순 변수 사용을 넘어 WHR×BMI 파생변수 생성, 비선형성을 반영한 연령 구간화(Binning), 다항항(Polynomial features) 추가로 변수의 설명력을 강화합니다.

데이터 불균형 대응

당뇨 발병 케이스 부족 문제를 해결하기 위해 Class Weight 조정, Focal Loss 함수 적용, 또는 예측 임계값(Threshold) 최적화를 통해 Recall 성능을 개선합니다.

교차 검증 전략 강화

과적합 방지를 위해 Nested CV를 도입하고, Repeated Stratified K-Fold로 평가 신뢰도를 확보하여 일반화 성능을 극대화합니다.

앙상블 및 보정

단일 모델 한계를 극복하기 위해 LGBM, XGBoost 모델 스태킹(Stacking) 및 Platt Scaling 등을 통한 예측 확률 보정(Calibration)을 수행합니다.

모델 해석력 확보

블랙박스 모델의 투명성을 높이기 위해 SHAP 값 분석 및 부분 의존성(Partial Dependence) 플롯을 통해 변수 간 상호작용을 시각적으로 규명합니다.

제출 전략 최적화

테스트 데이터에 대한 신뢰도 높은 예측값을 다시 학습에 사용하는 Pseudo-labeling 기법 적용 및 Public LB 점수 추적을 통한 사후 튜닝을 진행합니다.