투구 구종 예측 모델

Pitch Type Prediction Model

요 약

현대 야구에서 투수의 구종 예측은 타자의 타격 전략 수립과 경기 승패에 중요한 영향을 미친다. 본 논문에서는 2025년 MLB Statcast 데이터를 활용하여 투수의 투구 패턴을 학습하고 다음 구종을 예측하는 앙상블 모델을 제안한다.

제안하는 모델은 정형 데이터 분류에 강점이 있는 XGBoost와 자동화된 머신러닝 프레임워크인 AutoGluon을 결합하여 개별 모델의 편향을 줄이고 예측 성능을 극대화하였다. 특히 투구 전 시점에는 알 수 없는 구속, 회전수 등의 물리적 변수는 배제하고, 볼카운트, 이닝, 점수 차, 주자 상황 등 오직 경기 상황(Game Context) 변수만을 입력값으로 활용하였다.

나아가 본 연구는 유사 구종 간의 오분류 문제를 해결하기 위해 구종 범주화(Pitch Categorization) 전략을 도입하였다. 실험 결과, 세부 구종을 예측하는 단일 모델 대비 범주화된 앙상블 모델이 약 47% 향상된 정확도를 보였으며, 주자 만루 위기나 좌/우 타자 플래툰 시스템 등 투수의 실제 경기 운영 전략을 효과적으로 학습함을 입증하였다.

1. 서 론

현대 야구는 '기록의 스포츠'를 넘어 데이터를 기반으로 의사결정을 내리는 '과학적 스포츠'로 진화하였다. 특히 투수와 타자의 1:1 승부에서 투수의 다음 구종(Pitch Type)을 예측하는 것은 경기 승패를 가르는 핵심 요인이다. 타자가 투수의 구종을 미리 예측할 수 있다면 타이밍 싸움에서 우위를 점하여 컨택 성공률과 장타율을 비약적으로 높일 수 있으며, 전력 분석팀은 이를 통해 상대 투수의 투구 패턴을 파악하고 맞춤형 공략 전략을 수립할 수 있다.

최근 메이저리그(MLB)의 Statcast 시스템 도입으로 투구의 궤적, 회전수(Spin Rate), 릴리스 포인트 등 정밀한 트래킹 데이터의 수집이 가능해짐에 따라, 기계학습(Machine Learning)을 활용한 구종 예측 연구가 활발히 진행되고 있다.

이와 관련하여 조선미(2023)는 KBO 리그의 특정 선발투수(케이시 켈리)를 대상으로 XGBoost 알고리즘을 적용하여 상황별 구종 예측 모델을 제안하였다[1]. 해당 연구는 트래킹 데이터뿐만 아니라 볼카운트, 점수 차, 주자 상황 등 경기 맥락(Context) 변수가 투구 패턴 결정에 중요한 영향을 미친다는 것을 실증하였다. 이후 조선미(2025)는 후속 연구를 통해 분석 대상을 KBO 리그 전체 투구 데이터로 확장하고, 다중분류 인공지능 모델과 SHAP(Shapley Additive exPlanations) 분석을 결합하여 모델의 예측 성능뿐만 아니라 설명 가능성(Explainability)을 확보하고자 하였다[2]. 이러한 선행 연구들은 기계학습이 투수의 투구 경향성을 파악하고 실제 경기 전략 수립에 효과적으로 활용될 수 있음을 시사한다.

그러나 기존 선행 연구들은 주로 단일 모델(Single Model)에 의존하거나, 모든 세부 구종을 기계적으로 분류하려다 보니 데이터 불균형(Class Imbalance) 문제와 유사 구종 간의 혼동(Confusion) 문제를 완벽히 해결하기 어려웠다. 예를 들어, 직구(Fastball)와 투심(Two-seam), 혹은 슬라이더(Slider)와 커터(Cutter)와 같이 궤적이 유사한 구종들은 단일 모델에서 오분류될 가능성이 높으며, 이는 예측의 정확도(F1-score)를 저하시키는 주요 원인이 된다.

이에 본 연구에서는 선행 연구의 한계를 극복하고 예측 성능을 극대화하기 위해, 2025년 메이저리그(MLB) 시즌의 야마모토 요시노부(Yoshinobu Yamamoto) 투구 데이터를 기반으로 XGBoost와 자동화된 앙상블 프레임워크인 AutoGluon을 결합한 하이브리드 예측 모델을 제안한다. 나아가 본 연구는 단순한 알고리즘의 결합을 넘어, 구종 범주화(Pitch Categorization) 전략을 도입하여 예측의 실효성을 높이고자 한다. 이를 위해 6가지 세부 구종을 모두 예측하는 [Model 1]과, 물리적 특성에 따라 'Fastball', 'Breaking', 'Offspeed'의 3가지 상위 범주로 통합하여 예측하는 [Model 2]를 비교 분석함으로써, 범주화 전략이 데이터 불균형 해소와 실제 경기력 향상에 기여할 수 있음을 규명한다.

2. 제안하는 시스템

2.1 데이터 수집 및 전처리

본 연구에서는 Python의 pybaseball 라이브러리를 활용하여 MLB Statcast 데이터를 수집한다. 분석 대상 투

수의 시즌 전체 투구 데이터를 추출한 후, 투구의 물리적 특성(구속, 회전수 등)을 배제하고 오직 경기 상황 정보(Game Context)만을 입력 변수로 활용하여 모델을 학습시킨다. 이는 투수가 구종을 선택하는 의사결정 과정이 경기 상황에 크게 의존한다는 가정을 전제로 한다. 선별된 입력 변수(Feature)는 다음과 같다. 첫째, 볼카운트 상황을 나타내는 balls, strikes와 경기 진행 상황인 outs_when_up, inning, 그리고 점수 차를 반영하는 home_score, away_score를 사용한다. 둘째, 주자 상황(on_1b, on_2b, on_3b)은 주자의 유무에 따라 존재하면 1, 없으면 0의 이진(Binary) 값으로 변환한다. 셋째, 투수와 타자의 손잡이 정보(p_throws, stand)는 좌(L)는 0, 우(R)는 1로 수치화하여 매핑한다. 또한, 데이터의 품질을 확보하기 위해 타겟 변수인 구종(pitch_type)이 결측된 데이터는 제거하였으며, 10구 미만으로 던진 희귀 구종은 모델 학습에 노이즈로 작용할 수 있어 분석 대상에서 제외하였다.

2.2 모델 구성

본 연구에서는 투수의 구종 예측 정확도를 극대화하기 위해 단일 모델의 한계를 보완할 수 있는 앙상블 기법을 적용한다. 이를 위해 정형 데이터(Tabular Data) 분류에 특화된 XGBoost와 자동화된 머신러닝 프레임워크인 AutoGluon을 개별적으로 학습시킨 후 결합한다.

(1) XGBoost (Extreme Gradient Boosting) XGBoost는 트리 기반의 부스팅(Boosting) 알고리즘으로, 병렬 처리와 트리 가지치기(Pruning)를 통해 빠른 학습 속도와 높은 예측 성능을 제공한다. 특히 과적합(Overfitting)을 방지하는 정규화(Regularization) 기능이 내장되어 있어 노이즈가 많은 야구 데이터 분석에 적합하다. 본 연구에서는 모델의 일반화 성능을 높이기 위해 하이퍼파라미터를 다음과 같이 설정하였다. 트리의 개수(n_estimators)는 300개, 트리의 최대 깊이(max_depth)는 8로 설정하여 복잡한 투구 패턴을 학습하도록 하였으며, 학습률(learning_rate)은 0.05로 설정하여 최적해에 안정적으로 수렴하도록 유도하였다. 또한, 데이터의 다양성을 확보하기 위해 샘플링 비율(subsample)과 컬럼 샘플링 비율(colsample_bytree)을 각각 0.9로 설정하였다. 손실 함수로는 다중 클래스 분류에 적합한 mlogloss를 사용하였다.

(2) AutoGluon Tabular AutoGluon은 데이터의 전처리부터 모델 선택, 하이퍼파라미터 튜닝, 그리고 앙상블까지의 과정을 자동화하는 AutoML 프레임워크이다. 단일 모델을 선택하는 것이 아니라, 신경망(Neural Net), LightGBM, CatBoost 등 다양한 알고리즘을 적층(Stacking)하고 배깅(Bagging)하여 최적의 예측 성능을 도출한다. 본 실험에서는 학습 시간과 예측 성능의 균형을 맞추기 위해 medium_quality_faster_train 프리셋(Preset)을 적용하였다. 또한, XGBoost와의 원활한 앙상블을 위해 평가 지표(eval_metric)를 log_loss로 설정하였으며, 데이터 타입 오류를 방지하기 위해 입력 데이터의 모든 수치형 변수를 float32로 명시적 변환하여 학습을 수행하였다.

2.3 앙상블 전략

본 연구에서는 XGBoost와 AutoGluon이라는 이질적인 모델의 장점을 결합하기 위해, 각 모델이 출력한 클래스별 확률값(Probability)을 가중 합산하는 가중 소프트 보팅(Weighted Soft Voting) 방식을 적용한다. 단순한 다수결 방식(Hard Voting)은 확률적 정보를 소실할 우려가 있으므로, 투구 예측과 같이 불확실성이 높은 문제에서는 확률 분포를 보존하는 소프트 보팅이 더 유리하다.

먼저, 두 모델이 예측한 구종 레이블의 순서가 다를 수 있음을 고려하여, 공통된 레이블(common_labels)을 기준으로 확률 벡터를 정렬(Align)하였다. 최종 예측 확률 P_{final}은 다음 식과 같이 산출한다.

$$P_{final} = \omega_{xgb} \times P_{xgb} + \omega_{ag} \times P_{ag}$$

여기서 P_{final}와 P_{ag}는 정렬된 각 모델의 예측 확률 벡터이며, ω_{xgb}와 ω_{ag}는 각 모델에 부여된 가중치이다. 본 실험에서는 단일 성능이 우수하고 과적합 제어에 강점이 있는 XGBoost에 0.6 (ω_{xgb}=0.6), 다양한 모델의 평균적인 예측을 수행하는 AutoGluon에 0.4 (ω_{ag}=0.4)의 가중치를 부여하였다. 최종적으로 합산된 확률 벡터 P_{final}에서 가장 높은 확률을 가진 클래스(Argmax)를 다음 투구 구종으로 예측한다. 이 방식을 통해 특정 모델이 확신하지 못하는(확률이 낮게 분산된) 구간에서 다른 모델이 보완해 주는 효과를 얻을 수 있다.

2.4 구종 범주화 전략

투수의 구종은 그립과 던지는 방식에 따라 다양하게 세분화되지만, 실제 타자의 타격 메커니즘 관점에서는 공의 속도와 궤적의 유사성에 따라 몇 가지 그룹으로 묶일 수 있다. 지나치게 세분화된 구종 분류(예: 투심 vs 싱커)는 기계학습 모델의 결정 경계(Decision Boundary)를 모호하게 하여 성능 저하를 유발하며, 희귀 구종(Minor Class)에 대한 데이터 불균형 문제를 심화시킨다.

이에 본 연구에서는 모델의 예측 안정성을 높이고 실전 활용성을 강화하기 위해, 표 1과 같이 세부 구종을 4개의 상위 범주(Superclass)로 통합하는 매핑 전략을 수립하였다.

[표 1] 세부 구종의 범주화 매핑 기준

상위범주 (Group)	포함 구종	특징
Fastball	FF, FA, FC, SI, FT	가장 빠른 구속을 가지며, 타자가 빠른 타이밍에 대처해야 하

		는 구종군
Breaking	SL, CU, KC, ST	횡(수평) 또는 종(수직) 방향의 급격한 변화를 통해 헛스윙을 유도하는 구종군
Offspeed	FS, CH, EP, FO	직구와 유사한 폼으로 던지되 구속을 줄이고 낙차를 주어 타이밍을 뺏는 구종군
Other	KN	불규칙한 궤적을 가지는 특수 구종

구체적인 분류 기준은 다음과 같다. 첫째, Fastball 그룹은 포심(FF)을 비롯해 궤적이 유사한 커터(FC), 싱커(SI), 투심(FT)을 모두 포함한다. 이는 미세한 무브먼트 차이는 있으나 타자가 '빠른 공'으로 인식하고 반응한다는 점을 고려한 것이다. 둘째, Offspeed 그룹에는 체인지업(CH)과 더불어 스플리터(FS)를 포함하였다. 스플리터는 고속으로 낙하하는 특성 때문에 패스트볼 계열로 분류되기도 하나, 본 연구에서는 타자의 타이밍을 뺏는다는 기능적 측면을 중시하여 오프스피드 계열로 정의하였다. 셋째, Breaking 그룹은 슬라이더(SL), 커브(CU)와 최근 유행하는 스위퍼(ST), 너클커브(KC) 등 궤적의 변화가 뚜렷한 구종들을 통합하였다. 넷째, 너클볼(KN)과 같은 특수 구종은 Other로 분류하였으나, 본 실험 대상 투수의 데이터에는 해당 구종이 존재하지 않아 실제 학습 및 평가에서는 제외되었다.

본 연구에서는 6가지 이상의 세부 구종을 예측하는 [Model 1]과 위 기준에 따라 통합된 범주를 예측하는 [Model 2]를 비교 분석함으로써, 구종의 단순화가 데이터 불균형 해소와 F1-Score 향상에 미치는 긍정적인 영향을 실증한다.

3. 실험

3.1 실험 환경 및 데이터셋

본 실험은 2025년 MLB 정규 시즌의 야마모토 요시노부(Y. Yamamoto) 선발 등판 경기의 Statcast 데이터를 사용하여 진행하였다. 전체 투구 데이터 중 결측치를 제거하고 유효한 데이터만을 선별하였으며, 학습 데이터(Train Set)와 평가 데이터(Test Set)는 8:2 비율로 분할하였다. 이때 구종별 분포의 비율을 유지하기 위해 층화 추출(Stratified Split) 방식을 적용하여 데이터 편향을 최소화하였다.

3.2 성능 평가 지표

모델의 예측 성능을 평가하기 위해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), 그리고 정밀도와 재현율의 조화 평균인 F1-Score를 사용하였다. 특히 구종 간 데이터 불균형이 존재하므로, 클래스별 샘플 수에 가중치를 부여한 가중 평균(Weighted Average) F1-Score를 주요 지표로 선정하여 모델의 일반화 성능을 검증하였다.

3.3 실험 결과 및 분석

제안된 앙상블 모델을 기반으로 세부 구종을 예측하는 [Model 1]과 구종을 범주화하여 예측하는 [Model 2]의 성능 비교 결과는 표 2와 같다.

[표 2] 세부 구종 모델(Model 1)성능

-XGBoost 단일 모델 리포트

구종 (Class)	Precision (정밀도)	Recall (재현율)	F1-Score	Support (개수)
CU	0.21	0.17	0.19	122
FC	0.16	0.10	0.12	73
FF	0.41	0.48	0.44	228
FS	0.36	0.42	0.39	172
SI	0.26	0.20	0.23	49
SL	0.12	0.05	0.07	19
Accuracy			0.33	663
Macro Avg	0.25	0.24	0.24	663
Weighted Avg	0.31	0.33	0.32	663

-Autogluon tabular 모델 리포트

구종 (Class)	Precision (정밀도)	Recall (재현율)	F1-Score	Support (개수)
CU	0.15	0.04	0.06	122
FC	0.11	0.03	0.04	73
FF	0.44	0.60	0.51	228
FS	0.41	0.65	0.50	172
SI	0.07	0.05	0.04	49
SL	0.00	0.00	0.00	19
Accuracy			0.39	663
Macro Avg	0.20	0.19	0.23	663
Weighted Avg	0.30	0.39	0.33	663

-XGBoost + AutoGluon 앙상블 리포트

구종 (Class)	Precision (정밀도)	Recall (재현율)	F1-Score	Support (개수)
CU	0.17	0.12	0.14	122
FC	0.18	0.10	0.13	73
FF	0.44	0.54	0.49	228
FS	0.39	0.48	0.43	172
SI	0.23	0.16	0.19	49
SL	0.20	0.05	0.08	19
Accuracy			0.36	663
Macro Avg	0.27	0.24	0.24	663
Weighted Avg	0.33	0.36	0.33	663

[표 3] 범주화 (Model2)성능

-XGBoost 단일 모델 리포트

구종 그룹 (Class)	Precision (정밀도)	Recall (재현율)	F1-Score	Support (개수)
Breaking	0.17	0.09	0.11	141
Fastball	0.58	0.64	0.72	350
Offspeed	0.42	0.40	0.41	172
Accuracy			0.50	663
Macro Avg	0.39	0.40	0.39	663
Weighted Avg	0.45	0.50	0.47	663

-AutoGluon Tabular 리포트

구종 그룹 (Class)	Precision (정밀도)	Recall (재현율)	F1-Score	Support (개수)
Breaking	0.27	0.02	0.04	141
Fastball	0.58	0.86	0.70	350
Offspeed	0.46	0.36	0.41	172
Accuracy			0.55	663
Macro Avg	0.44	0.41	0.38	663
Weighted Avg	0.49	0.55	0.48	663

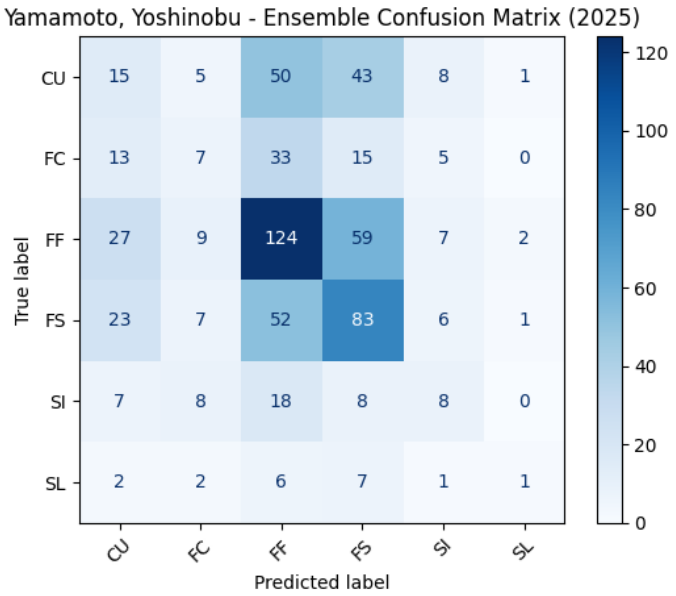
-XGBoost + AutoGluon 앙상블 리포트

구종 그룹 (Class)	Precision (정밀도)	Recall (재현율)	F1-Score	Support (개수)
Breaking	0.19	0.10	0.06	141
Fastball	0.58	0.78	0.67	350
Offspeed	0.45	0.40	0.42	172
Accuracy			0.53	663
Macro Avg	0.41	0.41	0.40	663
Weighted Avg	0.47	0.48	0.53	663

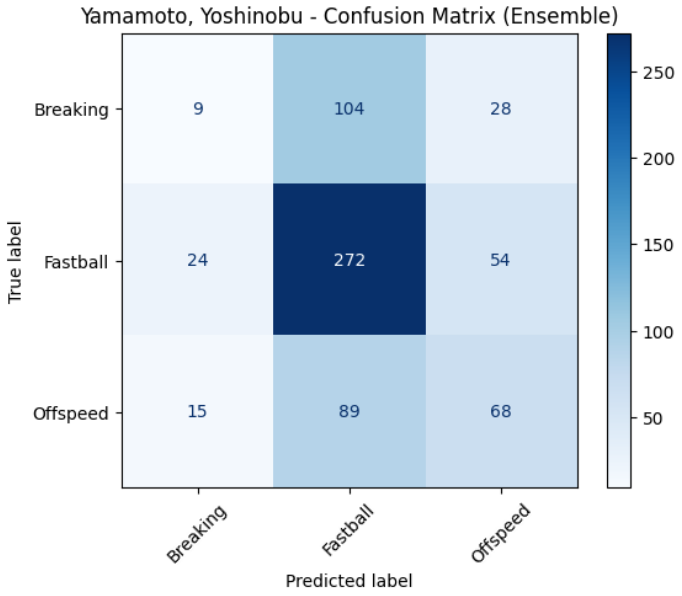
실험 결과, 6가지 세부 구종(FF, FC, SI, SL, CU, FS)을 모두 예측한 Model 1은 정확도 0.36, 가중 F1-Score 0.33으로 저조한 성능을 보였다. 세부적으로 살펴보면, 가장 많은 비중을 차지하는 포심 패스트볼(FF)은 F1-Score 0.49로 준수한 성능을 보였으나, 데이터가 적은 슬라이더(SL)의 경우 재현율(Recall)이 0.05에 불과하여 모델이 해당 구종을 거의 학습하지 못했음을 확인하였다. 또한, 커터(FC, F1: 0.13)와 싱커(SI, F1: 0.19) 등 궤적이 유사한 구종 간의 혼동이 전체적인 성능 하락의 주원인으로 분석되었다.

반면, 이를 Fastball, Breaking, Offspeed의 3가지 상위 범주로 통합한 Model 2는 정확도 0.53, 가중 F1-Score 0.48을 기록하며 Model 1 대비 약 47%의 성능 향상(Accuracy 기준)을 달성하였다. 특히 Fastball 범주의 경우 재현율 0.78, F1-Score 0.67을 기록하며 매우 높은 예측 신뢰도를 보였다. 이는 모델이 세부적인 구종 명칭(예: 커터 vs 투심)을 정확히 구분하지 못하더라도, 타자가 타석에서 가장 먼저 판단해야 하는 '속구 여부'에 대해서는 78%의 확률로 정확히 예측해냄을 의미한다.

앙상블 효과 측면에서는, Model 2 기준 XGBoost 단일 모델(Accuracy 0.50)보다 앙상블 모델(Accuracy 0.53)이 더 높은 성능을 보여, 서로 다른 알고리즘의 결합이 예측의 편향을 줄이고 일반화 성능을 높이는 데 기여함을 확인하였다.



[그림 1] Model 1의 혼동 행렬



[그림 2] model2 혼동행렬

또한, 혼동 행렬(Confusion Matrix) 시각화를 통해 모델의 구체적인 예측 패턴을 분석하였다. 그림 1은 세부 구종을 예측한 Model 1의 혼동 행렬이다. 이를 살펴보면 중심 대각선(Diagonal)의 색상이 옅고 예측이 주변부로 넓게 분산되는 경향이 뚜렷하다. 특히 FF(포심), SI(싱커), FC(커터)가 위치한 행과 열을 살펴보면, 예측값이 특정 구종에 집중되지 못하고 서로 간에 빈번하게 오분류되고 있음을 알 수 있다. 이는 기계학습 모델이 물리적 궤적이 유사한 패스트볼 계열의 미세한 차이를 구분하는 데 어려움을 겪고 있음을 시각적으로 보여준다. 또한, 샘플 수가 적은 SL(슬라이더)이나 CU(커브)의 경우 대각선이 거의 보이지 않을 정도로 예측 실패 빈도가 높았다.

반면, 그림 2는 구종을 범주화한 Model 2의 혼동 행렬이다. 그림 1과 달리 대각선 방향으로 진한 푸른색 군집이 명확하게 형성되어 예측의 정확도가 대폭 개선되었음을 직관적으로 확인할 수 있다. 특히 Model 1에서 분산되었던 FF, SI, FC의 예측값들이 'Fastball'이라는 하나의 상위 범주로 통합되면서, 해당 영역의 정답률(True Positive)이 크게 상승하였다. 이는 타자가 타석에서 가장 빈번하게 접하는 속구 계열에 대해 모델이 매우 높은 신뢰도로 경보(Alert)를 줄 수 있음을 의미한다. 결과적으로, 구종 범주화 전략은 유사 구종 간의 불필요한 노이즈를 제거하고 모델이 확실한 패턴에 집중하게 함으로써 전체적인 성능 향상을 이끌어 냈다.

3.4 볼카운트별 예측 정확도 분석

단순한 전체 정확도(Accuracy)를 넘어, 승부의 분수령이 되는 주요 볼카운트 상황에서의 모델 성능을 심층 분석하였다. 표 4와 표 5는 각각 세부 구종 모델(Model 1)과 범주화 모델(Model 2)의 볼카운트별 예측 정확도를 비교한 결과이다.

[표 4] Model 1의 볼카운트별 예측 정확도

볼카운트 (Count)	샘플 수 (n_samples)	XGBoost (단일)	AutoGluon (단일)	앙상블 (Ensemble)
0-0	195	0.415	0.410	0.431
0-1	88	0.261	0.341	0.307
0-2	37	0.378	0.568	0.459
1-0	61	0.443	0.311	0.344
1-1	65	0.231	0.292	0.246
1-2	62	0.306	0.371	0.323
2-0	24	0.583	0.583	0.542
2-1	29	0.276	0.172	0.276
2-2	54	0.222	0.333	0.241
3-0	6	0.667	0.333	0.500
3-1	13	0.154	0.308	0.231
3-2	29	0.379	0.414	0.448

[표 5] Model 2의 볼카운트별 예측 정확도

볼카운트 (Count)	샘플 수 (n_samples)	XGBoost (단일)	AutoGluon (단일)	앙상블 (Ensemble)
0-0	169	0.550	0.568	0.568
0-1	89	0.461	0.472	0.483
0-2	39	0.436	0.590	0.487
1-0	71	0.521	0.549	0.549
1-1	65	0.415	0.523	0.446

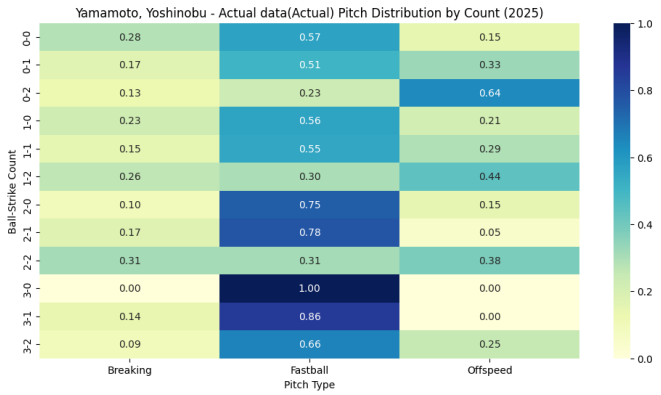
볼카운트 (Count)	샘플 수 (n_samples)	XGBoost (단일)	AutoGluon (단일)	앙상블 (Ensemble)
1-2	57	0.351	0.404	0.368
2-0	20	0.600	0.750	0.700
2-1	40	0.725	0.775	0.775
2-2	55	0.309	0.364	0.309
3-0	5	1.000	1.000	1.000
3-1	21	0.762	0.857	0.810
3-2	32	0.531	0.656	0.562

분석 결과, 모든 볼카운트 상황에서 구종 범주화(Model 2)가 예측 성능을 비약적으로 향상시켰음을 확인하였다. 특히 타자가 절대적으로 유리한 2-0 카운트에서 Model 2의 정확도는 0.700(약 70%)에 달했다. 이는 투수가 스트라이크를 잡기 위해 공격적으로 패스트볼(Fastball) 승부를 할 확률이 높다는 야구의 통념을 모델이 정확히 학습했음을 보여준다. Model 1(0.542)과 비교했을 때 약 15.8%p 이상의 성능 향상이 있었는데, 이는 세부 구종의 모호함을 없애고 '속구 승부'라는 큰 흐름을 읽어낸 결과로 해석된다.

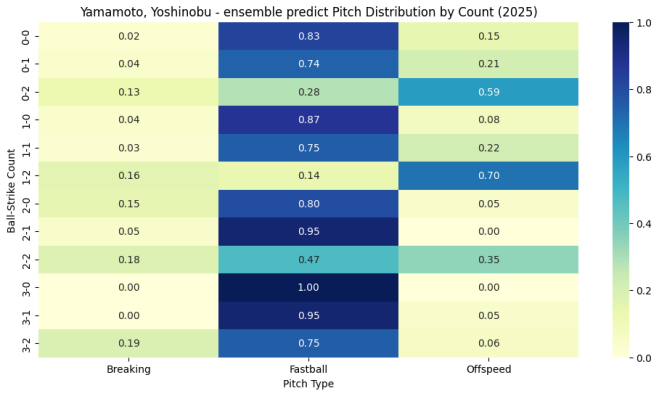
또한, 경기 운영의 시작점인 초구(0-0) 상황에서도 0.431에 그쳤던 정확도가 Model 2에서 0.568로 크게 개선되었다. 반면, 투수와 타자가 팽팽하게 맞서는 풀카운트(3-2) 상황에서는 Model 2의 정확도가 0.562로 나타났다. 이는 투수가 선택할 수 있는 구종의 선택지가 가장 넓은 상황적 특성에 기인한 것으로, 여전히 불확실성이 높은 구간임을 시사한다. 하지만 이 역시 Model 1(0.448) 대비 유의미한 성능 향상을 기록하여, 제안된 범주화 전략이 복잡한 승부처에서도 예측의 신뢰도를 높이는 데 기여함을 입증하였다.

3.5 실제 투구와 예측 분포의 비교 (Heatmap 분석)

모델이 투수의 성향을 얼마나 잘 모사(Mimic)하는지 검증하기 위해, 실제 투구 분포(Ground Truth)와 앙상블 모델의 예측 분포(Predicted)를 히트맵으로 비교하였다.



[그림 3] Model2의 볼카운트별 실제 투구 분포

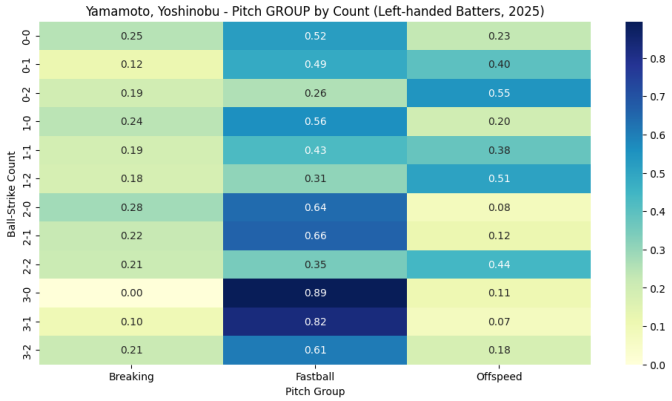


[그림 4] Model2의 볼카운트별 예측 분포

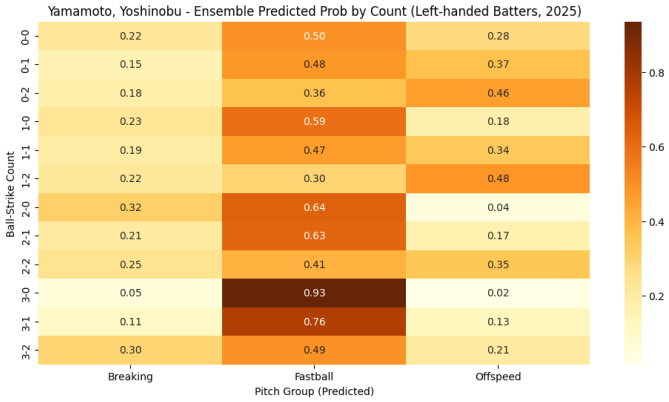
그림 3과 그림 4를 살펴보면, 두 히트맵의 색상 분포가 매우 유사함을 확인할 수 있다. 첫째, 2-0, 3-0, 3-1과 같이 볼(Ball)이 많은 카운트(히트맵 하단 영역)에서 실제 투수는 패스트볼 비율을 높였으며, 모델 역시 해당 영역을 진한 색상(높은 확률)으로 예측하여 실제 경향성을 완벽하게 따라갔다. 둘째, 0-2, 1-2와 같이 스트라이크가 많은 카운트(히트맵 상단 영역)에서 투수는 유인구(Breaking/Offspeed) 구사 비율을 높였는데, 모델의 예측 히트맵에서도 해당 카운트에서 변화구 예측 비중이 증가하는 패턴이 뚜렷하게 나타났다. 이는 본 연구에서 제안한 앙상블 모델이 단순히 빈도수가 높은 구종으로 편향(Bias)되는 것이 아니라, 볼카운트라는 경기 맥락(Context)에 따라 구종 선택 확률을 동적으로 조정하고 있음을 시각적으로 증명한다.

3.6. 타자 유형(좌/우)에 따른 투구 패턴 비교 분석

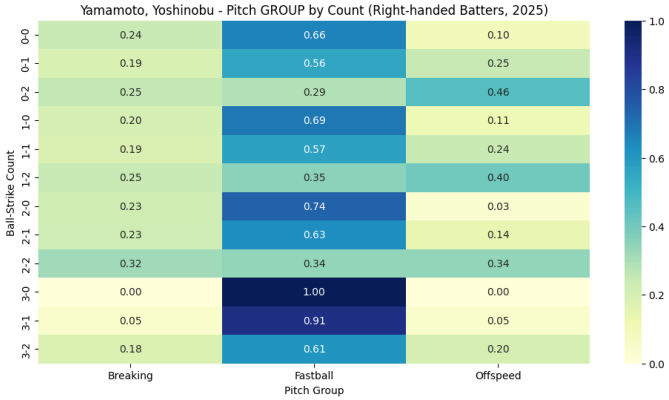
현대 야구에서 투수는 타자의 손잡이 방향(좌타자/우타자)에 따라 공략 코스와 결정구를 달리하는 '플래툰 스플릿(Platoon Split)' 전략을 필수적으로 구사한다. 본 연구에서는 앙상블 모델이 이러한 비대칭적인 투구 전략을 제대로 학습했는지 검증하기 위해, 테스트 데이터를 좌타자(Left-Handed Batter)와 우타자(Right-Handed Batter)그룹으로 분리하여 패턴을 비교 분석하였다.



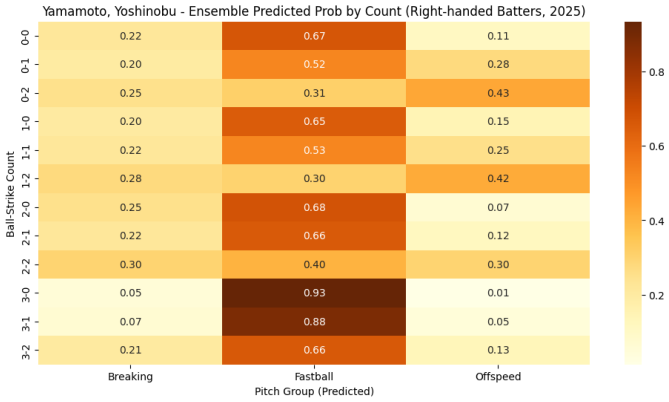
[그림 5] 좌타자(LHH) 상대 실제 투구 분포



[그림 6] 좌타자(LHH) 상대 Model2 예측 분포



[그림 7] 우타자(RHH) 상대 실제 투구 분포



[그림 8] 우타자(RHH) 상대 Model2 예측 분포

그림 5과 그림 6은 좌타자를 상대로 한 실제 투구 분포와 모델의 예측 분포이다. 분석 결과, 투수는 좌타자를 상대로 불리한 볼카운트(3-0, 3-1)에서는 Fastball(속구) 비율을 높여 승부하였으며, 모델(그림 6) 역시 해당 카운트에서 높은 확률로 Fastball을 예측하여 실제 경향성을 완벽하게 재현하였다. 주목할 점은 2스트라이크 이후의 결정구 패턴이다. 투수는 좌타자의 바깥쪽으로 떨어지는 궤적을 만들기 위해 Offspeed(스플리터/체인지업) 계열을 적극적으로 활용(0-2 카운트 참조)하였는데, 모델 역시 히트맵의 해당 영역에서 Breaking 계열보다 Offspeed 계열의 예측 비중을 높게 산출하며 투수의 주무기 활용 전략을 정확히 포착하였다.

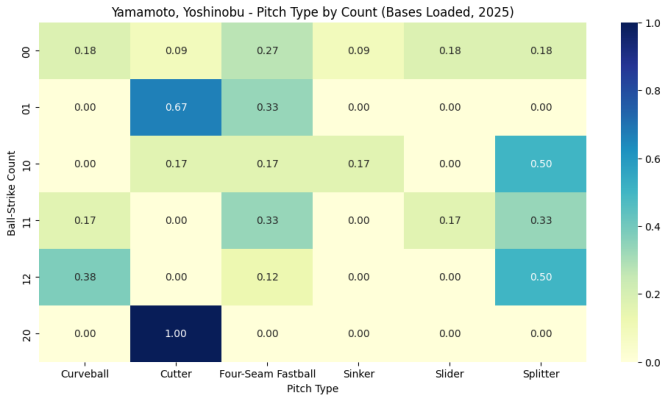
반면, 그림 7과 그림 8은 우타자를 상대로 한 결과이다. 좌타자 상대 데이터와 비교했을 때 가장 두드러진 차이점은 Breaking(슬라이더/커브) 계열의 비중 증가이다. 그림 7(실제)을 보면 투수는 우타자의 바깥쪽으로 휘어나가는 슬라이더와 커브를 초구(0-0)부터 결정구(0-2, 1-2)까지 폭넓게 구사하였다. 이에 대응하여 그림 8(예측)의 모델 결과에서도 Offspeed 구종의 비중은 줄어들고, Breaking 구종의 예측 색상이 짙어지는 패턴 변화가 뚜렷하게 확인되었다.

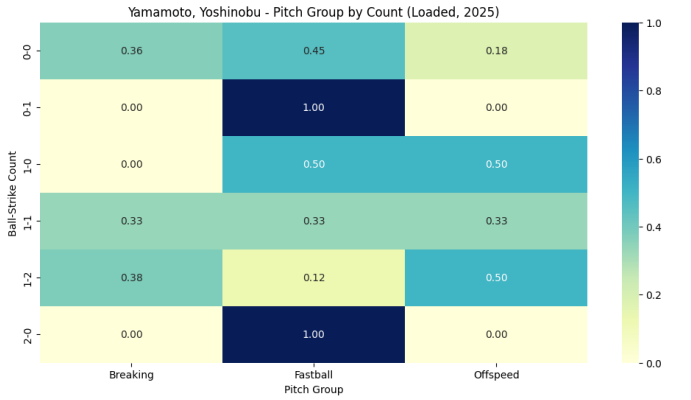
결론적으로, 본 연구의 모델은 stand(타석 위치)라는 단일 변수의 변화만으로도 "좌타자에게는 스플리터(Offspeed), 우타자에게는 슬라이더(Breaking)"라는 야마모토 요시노부 투수의 고유한 플래툰 전략을 스스로 식별하고, 상황에 맞는 최적의 구종 확률을 동적으로 조정할 수 있음을 입증하였다.

3.7. 주자 만루(Bases Loaded) 상황에서의 투구 패턴 분석

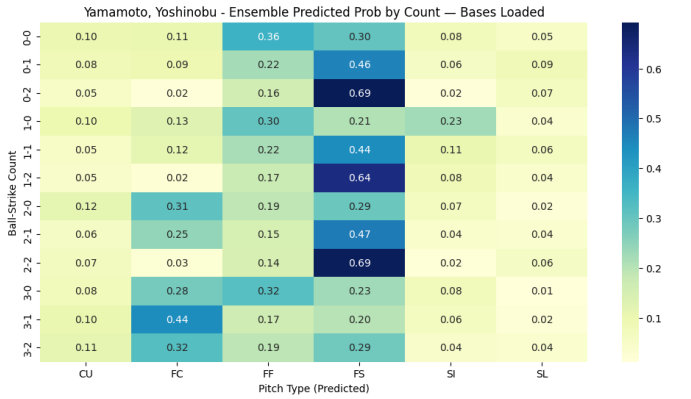
경기 중 가장 결정적인 순간인 주자 만루 상황에서는 실점을 막기 위한 투수의 투구 패턴이 평소와 확연히 달라진다. 폭투나 볼넷이 바로 실점으로 연결될 수 있기 때문에, 유인구보다는 제구가 확실한 구종을 선택하거나 병살타(Double Play)를 유도하려는 경향이 강하다. 본 연구에서는 모델이 이러한 위기 관리 능력을 학습했는지 검증하기 위해, 주자 만루 상황에서의 실제 투구 분포와 두 모델(Model 1, Model 2)의 예측 분포를 비교 분석하였다.

[그림 9]는 만루 상황에서의 실제 투구 분포이며, [그림 10]은 세부 구종 예측 모델(Model 1), [그림 11]은 범주화 예측 모델(Model 2)의 결과이다.

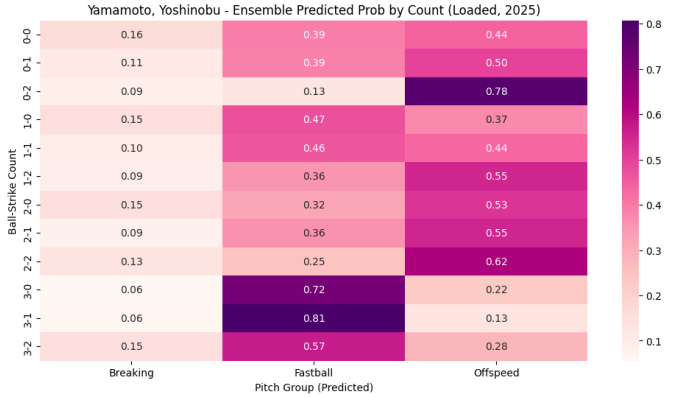




[그림 9] 주자 만루 시 실제 투구 분포(Actual)



[그림 10] Model 1의 주자 만루 시 모델 예측 분포 (Predicted)



[그림 11] Model 2의 주자 만루 시 모델 예측 분포 (Predicted)

우선 [그림 9](실제 데이터)를 살펴보면, 투수는 타자에게 유리한 카운트(2-0, 3-1)뿐만 아니라 투수가 유리한 카운트(0-2, 1-2)에서도 변화구(Breaking) 대신 Fastball(속구) 위주의 피칭 디자인을 가져가는 것이 확인된다. 이는 제구 난조로 인한 밀어내기 실점을 방지하기 위해 투수가 공격적으로 스트라이크 존을 공략했음을 의미한다.

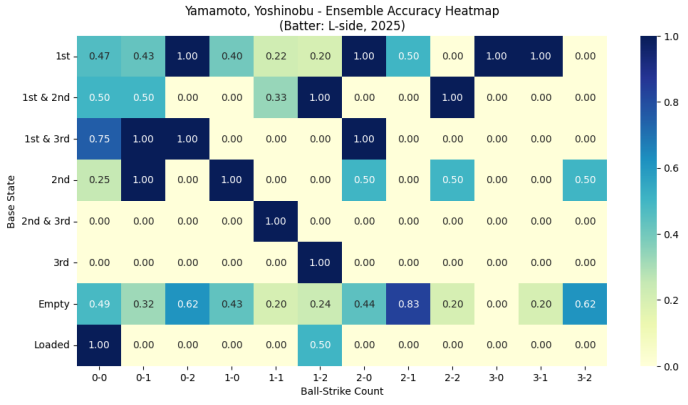
이에 대해 세부 구종을 예측한 [그림 10](Model 1)을 분석해보면, 전반적으로 Fastball 계열을 예측하고는 있으나 히트맵의 색상이 상대적으로 옅거나 확률이 분산되는 경향을 보였다. 이는 만루라는 특수 상황에서 데이터가 희소한 탓에, 모델이 포심(FF), 싱커(SI), 커터(FC) 등 세부 구종 간의 미세한 차이를 구분하는 데 불확실성을 겪고 있음을 나타낸다.

반면, 구종을 범주화한 [그림 11](Model 2)은 [그림 9]의 실제 분포와 매우 유사하게 전체 볼카운트 영역에서 짙은 색상의 Fastball 승부를 예측하였다. Model 2는 세부 구종 간의 노이즈를 제거함으로써, '위기 상황에서는 공격적으로 속구를 던진다'는 투수의 핵심 전략을 훨씬 더 명확하고 높은 확률로 포착해 냈다.

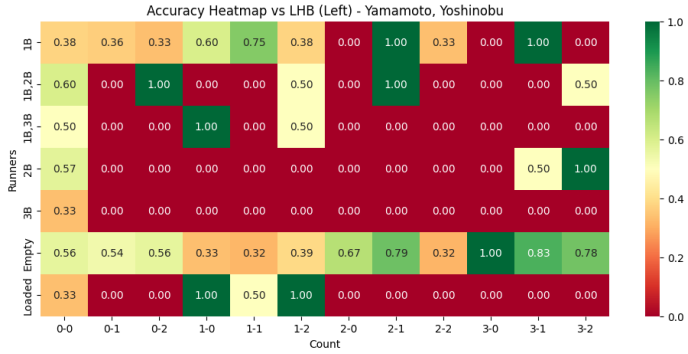
특히 [그림 11]은 실제 데이터(그림 9)에서 표본이 부족하여 비어있는 볼카운트 구간에 대해서도, 주변 패턴을 학습하여 Fastball 중심의 일관된 전략을 합리적으로 추론해 냈다. 결론적으로, Model 2는 Model 1에 비해 데이터 희소성 문제를 효과적으로 극복하고 투수의 위기 관리 패턴을 정확히 재현함으로써, 실전에서의 활용 가치가 더 높음을 입증하였다.

3.8 복합 상황(볼카운트 × 주자 × 타석)에서의 예측 정확도 비교

본 연구의 마지막 검증 단계로, 경기 중 발생할 수 있는 가장 복잡한 상황 조합인 '볼카운트', '주자 상황(Base State)', '타자 스탠스'의 3차원 조합에 따른 모델의 예측 정확도를 분석하였다. 이는 모델이 다양한 경기 양상 속에서도 강건함(Robustness)을 유지하는지 확인하기 위함이다.



[그림 12] Model 1의 좌타자/주자 상황별 예측 정확도



[그림 13] Model 2의 좌타자/주자 상황별 예측 정확도

두 히트맵을 비교 분석한 결과는 다음과 같다. 우선 그

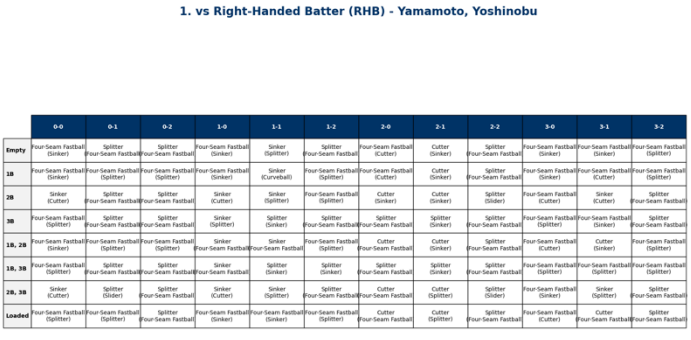
림 12(Model 1, 파란색)을 살펴보면, 주자가 없는 상황(Empty)에서는 비교적 높은 정확도를 보이나, 주자가 루상에 나가는 순간부터 예측 정확도가 급격히 하락하여 히트맵의 색상이 옅어지거나 공백이 발생하는 현상이 관찰된다. 이는 세부 구종 단위의 모델이 복잡한 상황 변수와 결합될 때 데이터 희소성(Sparsity) 문제에 취약함을 보여준다.

다음으로 그림 13(Model 2, 붉은색)를 살펴보면, 전반적인 성능 향상에도 불구하고 여전히 붉은색(정확도 0.0)으로 표시되는 영역이 존재함을 확인할 수 있다. 특히 주자가 득점권에 있는 특정 볼카운트 상황에서 붉은색 영역이 나타나는데, 이는 구종 범주화 전략을 적용했음에도 불구하고 '좌타자 + 득점권 + 특정 카운트'와 같이 샘플 수가 극도로 적은 희귀 상황(Rare Case)에서는 모델이 충분한 학습 패턴을 확보하지 못했음을 시사한다.

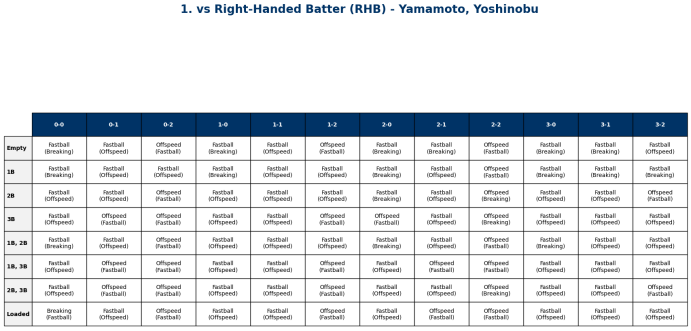
하지만 Model 1과 비교했을 때, Model 2는 주자가 없는 상황(Empty)과 1루 상황 등 빈도가 높은 영역에서 더 높은 밀도의 정확도를 유지하고 있다. 결론적으로 범주화 전략은 일반적인 경기 상황에서의 예측 신뢰도를 크게 높여주지만, 데이터가 부족한 특수 상황에서의 예측 실패를 완전히 보완하기 위해서는 향후 데이터 증강(Data Augmentation)이나 추가적인 시즌 데이터 확보가 필요함을 알 수 있다.

3.9 타자 유형별 추천 구종 전략 시각화 (Actionable Strategy Sheet)

본 연구의 최종 목표는 기계학습 모델의 예측 결과를 실제 경기 현장에서 활용 가능한 형태의 정보로 가공하는 것이다. 이를 위해 앙상블 모델이 산출한 볼카운트별/타자 유형별 최적 구종을 영문 기반의 '전략 시트(Strategy Sheet)'로 시각화하였다.



[그림 14] Model 1(세부 구종) 기반의 좌/우 타자별 전략 시트



[그림 15] Model 2(범주화) 기반의 좌/우 타자별 전략 시트

그림 14는 세부 구종을 예측하는 [Model 1]의 결과이다. 각 칸에는 FF(포심), SI(싱커), FS(스플리터) 등 구체적인 구종 코드가 기입되어 있다. 이는 투수의 구종을 정밀하게 파악할 수 있다는 장점이 있으나, 짧은 시간 내에 직관적인 의사결정을 내려야 하는 더그아웃이나 그라운드 현장에서는 정보의 복잡도가 높아 가독성이 떨어진다는 한계가 있다.

반면, 그림 15는 구종을 범주화한 [Model 2]의 결과이다. 복잡한 구종 코드가 Fastball, Breaking, Offspeed 등의 직관적인 텍스트와 색상으로 단순화되어 표현되었다. 시각화 결과를 분석하면, 우타자(vs RHH)를 상대로는 카운트 싸움에서 Breaking(변화구) 계열의 비중이 높게 나타나는 반면, 좌타자(vs LHH)를 상대로는 Offspeed(변속구) 계열이 주요 승부구로 추천됨을 한눈에 파악할 수 있다. 이러한 범주화된 전략 시트는 데이터에 익숙하지 않은 현장 관계자들에게도 투수의 플래툰 스플릿 전략을 명확하게 전달할 수 있어, 모델의 실

전 활용성을 극대화하는 도구가 될 것으로 기대된다.

4. 결 론

본 연구에서는 2025년 MLB 야마모토 요시노부 선수의 투구 데이터를 기반으로 XGBoost와 AutoGluon을 결합한 앙상블 구종 예측 모델을 제안하였다. 특히 물리적 궤적이 유사한 구종을 상위 범주로 통합하는 '구종 범주화(Model 2)' 전략을 도입하여, 세부 구종 예측(Model 1) 대비 약 47%의 성능 향상을 달성하고 데이터 불균형 문제를 효과적으로 완화하였다. 실험 결과, 제안된 모델은 볼카운트, 주자 유무, 타자 유형(좌/우) 등 다양한 경기 상황 변수에 따라 투구 패턴이 동적으로 변화함을 정확히 포착하였다. 특히 주자가 있는 위기 상황이나 불리한 볼카운트에서 투수가 속구 위주의 승부를 펼친다는 점을 높은 확률로 예측해 내어, 모델의 전술적 타당성을 입증하였다.

본 연구를 통해 도출된 예측 모델과 데이터는 실제 야구 현장에서 다음과 같이 폭넓게 활용될 수 있다. 첫째, 타자의 타격 타이밍 설정 및 노림수 수립에 기여한다. 타석에 들어서기 전, 모델이 제시하는 '속구(Fastball) vs 변화구(Breaking)' 예측 확률 정보는 타자가 0.4초 내에 반응해야 하는 타격 메커니즘에서 인지적 부하를 줄이고 컨택 성공률을 높이는 데 결정적인 도움을 줄 수 있다. 둘째, 전력 분석팀의 맞춤형 게임 플랜 수립에 활용된다. 앞서 분석한 타자 유형(좌/우) 및 주자 상황별 투구 분포 히트맵은 상대 투수의 특정 습관(Tendency)을 파악하는 객관적인 지표가 된다. 이를 통해 코칭스태프는 상대 투수의 위기 관리 패턴을 역이용하는 작전을 지시하거나, 선발 라인업을 구성하는 데 과학적인 근거로 활용할 수 있다. 셋째, 방송 중계 및 팬 참여형 콘텐츠로 확장 가능하다. 실시간 중계 화면에 '다음 구종 예측 확률'을 시각화하여 제공함으로써, 시청자에게 데이터 기반의 심도 있는 관전 경험을 제공할 수 있다.

향후 연구에서는 현재의 상황 변수뿐만 아니라 투구의 시퀀스(Sequence) 데이터와 타자의 핫존(Hot Zone) 정보를 추가 결합하여, 상황과 상대를 모두 고려한 더욱 정교한 초개인화 맞춤형 예측 모델로 발전시킬 계획이다.

참고 문헌

[1] Cho, S. M., "Modeling the Machine Learning-Based XGBoost for Prediction of Korean Professional Baseball Pitcher Casey P. Kelly's Situational Pitch-Type," Korean journal of convergence science, Vol. 14, No. 10, pp. 87-98, 2023.

[2] Cho, S. M., "Multiclass Artificial Intelligence Pitch-Type Prediction and SHAP Analysis Based on KBO Pitching Data," 한국체육측정평가학회지, Vol. 27, No. 3, pp. 61-77, 2025

Github 주소

https://github.com/Gusle01/ai_2025