AI기반 타자 성적 예측

딥러닝 및 LLM을 이용한 야구 성적 분석

2026-01-12

목차

프로젝트 개요 및 목표

프로젝트 개요

2025년 프로야구는 누적 관중 수 1200만 명을 돌파하며 높은 인기를 끌고 있다. 평소 선수 기록과 데이터 분석을 즐겨 보던 개인적인 관심을 계기로, 단순 기록 조회를 넘어 데이터 기반으로 선수의 미래 성적을 예측할 수 있는 시스템을 직접 구현해보고자 본 프로젝트를 기획하였다.

프로젝트 목표

본 프로젝트는 KBO 타자 데이터를 활용해 선수의 연령(에이징 커브)과 과거 시즌 성적 추이를 반영하여 다음 시즌 성적 증감을 예측하고, 그 결과를 Streamlit 기반 웹 서비스로 제공하며 LLM이 예측에 영향을 준 주요 요인(최근 성적, 에이징 커브 등)을 근거로 증감 **가능성**을 설명한다.

모델 개발 과정

데이터셋 구성

KBO 리그 타자 시즌별 성적 데이터

(원천 데이터: 1982–2025, 모델 학습 사용 구간: 2015년 이후)

출처:KBO player dataset.csv (kaggle)

데이터 전처리

대상- **2000년대 이후 데뷔한 KBO 리그 타자, 223타석 이상**

근거: 투수는 고려해야 할 변수가 많아 특징 정의가 복잡하므로, 타자 성적 예측에 집중했다. 또한 현역 선수중 최고령인 최형우 선수 (2004년 데뷔)를 감안하여 2000년대 이후 데뷔한 타자들을 대상으로 선정하였다.

단위- **선수 × 시즌**

기간- **2015년 이후 시즌**

근거:현재 144경기 체제가 적용되기 시작한 2015년도 이후 출장한 타자에 한하여 기간을 정하였다.

주요 입력 변수: **나이, 타석 수(PA), 타율, 출루율, 장타율, OPS, 홈런, wrc+, war 등 성적 지표**

예측 대상(Target):**다음 시즌 성적 증감(Δ = next − current)**

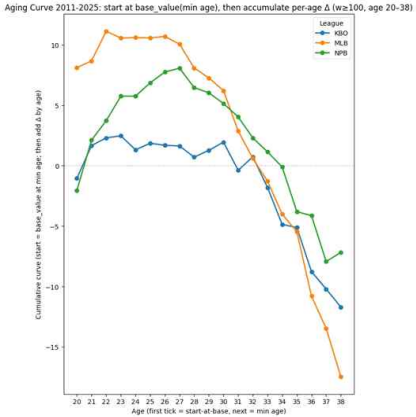
epoch:120

```
200
WAR       | MAE: 1.2481 | RMSE: 1.6794
wRC_plus  | MAE: 17.3149 | RMSE: 21.2696

300
WAR       | MAE: 1.2181 | RMSE: 1.6605
wRC_plus  | MAE: 15.0032 | RMSE: 19.5118

446
WAR       | MAE: 1.2468 | RMSE: 1.6057
wRC_plus  | MAE: 15.9058 | RMSE: 19.6547
```

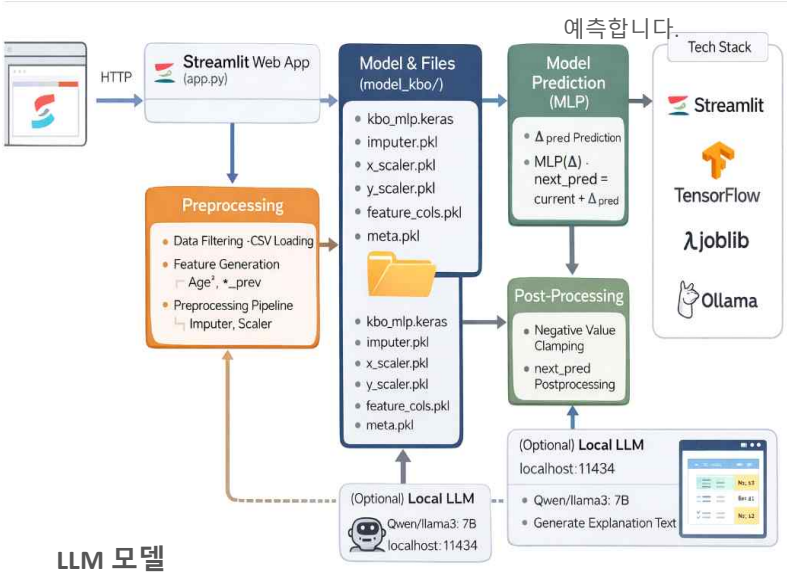
타석이 증가할수록 MAE와 RMSE는 감소하였지만 좋은 현상이라고만 볼 수는 없다고 판단하였다. 타석수가 너무 적으면 성적의 변동 폭이 과다할 수 있고, 타석수가 많을 경우 상위타자 위주의 안정적인 데이터만 남기 때문에 타석수를 규정타석의 절반인 223타석으로 선정하였다.



실제 에이징 커브는 선형이 아닌 곡선 형태를 보인다. 특히 고령 구간에서 성적 하락이 가속되기 때문에 이를 반영하기 위해 Age^2 독립변수를 추가하여 함께 고려하도록 하였다.

모델 개발 과정

시스템 아키텍쳐



LLM 모델

로컬 LLM인 Qwen을 Ollama API를 통해 호출하여 성적 예측의 근거를 자연어로 해설합니다.

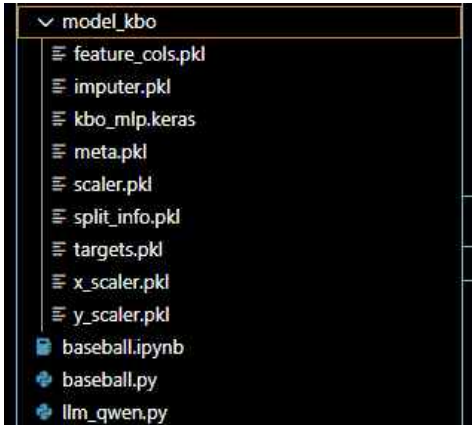
예측 모델

MLP기반 딥러닝 회귀 모델로 타자의 시즌별 수치 지표와 파생변수를 입력받아 다음 시즌 성적을 예측합니다.

웹 구현

Streamlit 기반 웹 어플리케이션을 통해 서버에서 데이터 로드->전처리->MLP예측->후처리->결과 시각화를 한 번에 수행해 화면에 출력합니다.

디렉토리 구조



model_kbo - 학습된 모델

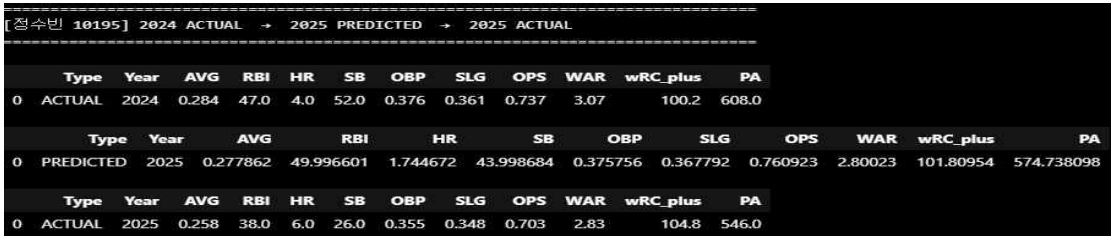
baseball.ipynb - 모델 학습/실험 노트북

baseball.py - 실행 파일

llm_qwen.py- llm모듈

기능 구현 - 예측 테스트

테스트 1



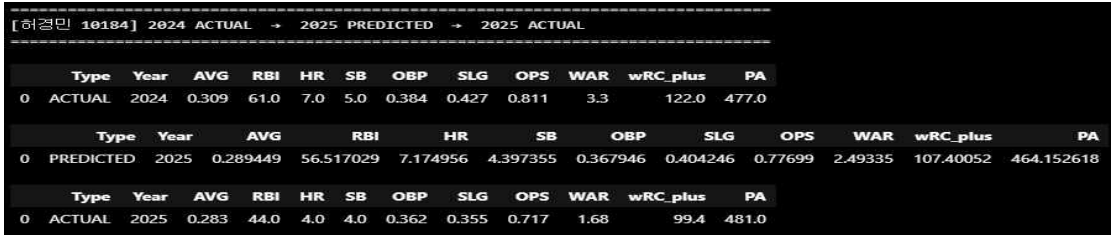
[정수빈 10195] 2024 ACTUAL → 2025 PREDICTED → 2025 ACTUAL

	Type	Year	AVG	RBI	HR	SB	OBP	SLG	OPS	WAR	wRC_plus	PA
0	ACTUAL	2024	0.284	47.0	4.0	52.0	0.376	0.361	0.737	3.07	100.2	608.0

	Type	Year	AVG	RBI	HR	SB	OBP	SLG	OPS	WAR	wRC_plus	PA
0	PREDICTED	2025	0.277862	49.996601	1.744672	43.998684	0.375756	0.367792	0.760923	2.80023	101.80954	574.738098

	Type	Year	AVG	RBI	HR	SB	OBP	SLG	OPS	WAR	wRC_plus	PA
0	ACTUAL	2025	0.258	38.0	6.0	26.0	0.355	0.348	0.703	2.83	104.8	546.0

테스트 2



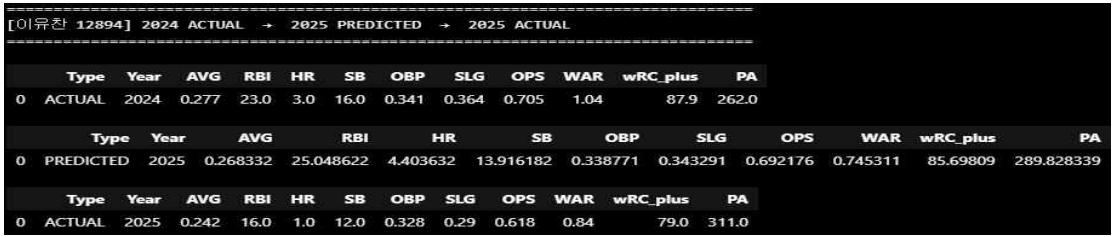
[허경민 10184] 2024 ACTUAL → 2025 PREDICTED → 2025 ACTUAL

	Type	Year	AVG	RBI	HR	SB	OBP	SLG	OPS	WAR	wRC_plus	PA
0	ACTUAL	2024	0.309	61.0	7.0	5.0	0.384	0.427	0.811	3.3	122.0	477.0

	Type	Year	AVG	RBI	HR	SB	OBP	SLG	OPS	WAR	wRC_plus	PA
0	PREDICTED	2025	0.289449	56.517029	7.174956	4.397355	0.367946	0.404246	0.77699	2.49335	107.40052	464.152618

	Type	Year	AVG	RBI	HR	SB	OBP	SLG	OPS	WAR	wRC_plus	PA
0	ACTUAL	2025	0.283	44.0	4.0	4.0	0.362	0.355	0.717	1.68	99.4	481.0

테스트 3



[이유찬 12894] 2024 ACTUAL → 2025 PREDICTED → 2025 ACTUAL

	Type	Year	AVG	RBI	HR	SB	OBP	SLG	OPS	WAR	wRC_plus	PA
0	ACTUAL	2024	0.277	23.0	3.0	16.0	0.341	0.364	0.705	1.04	87.9	262.0

	Type	Year	AVG	RBI	HR	SB	OBP	SLG	OPS	WAR	wRC_plus	PA
0	PREDICTED	2025	0.268332	25.048622	4.403632	13.916182	0.338771	0.343291	0.692176	0.745311	85.69809	289.828339

	Type	Year	AVG	RBI	HR	SB	OBP	SLG	OPS	WAR	wRC_plus	PA
0	ACTUAL	2025	0.242	16.0	1.0	12.0	0.328	0.29	0.618	0.84	79.0	311.0

테스트 결과

·2024년 타고투저로 성적이 높게 나온 상황에서도 모델은 최근 성적 흐름 + 나이(Age²) 를 반영해 2025년을 과대추정하지 않고 합리적으로 예측했다.

·실제 2025년 성적과 비교했을 때도 유지/하락과 같은 방향성(추세)을 비교적 잘 포착하는 것을 확인했다.

LLM 프롬프트

```
[출력 형식]
- 한국어 8~12문장
- 문장형 서술로 작성(표/코드/불릿 금지)
- 반드시 포함할 것:
  1) 한 줄 결론: "다음 시즌은 전년 대비 oo 방향(상승/하락/유지)으로 예측" (OPS/WAR 중심)
  2) 전년도(base_year)의 현재 스탯과 예측(next), 그리고 Δ(증감) 연결 설명
  3) 커리어 흐름(최근 N년): 최근 2~3년의 변화가 예측에 어떻게 반영됐는지
  4) 동나이대 비교: 동나이대(±1세) 집단 평균/중앙값과 비교해서 "평균 대비 어떤 편인지"
  5) PA(타석) 변화가 예측 해석에 주는 의미(표본/출전기회 관점) – 단, 근거 JSON 수치로만
  6) 마지막 문장에 "참고용" 안내

[해설 규칙]
- OPS, WAR, wRC+를 우선으로 설명하고, HR/RBI/AVG는 보조로 사용
- "동나이대 평균적인 aging trend(증감)"과 "선수 개인 커리어 trend(증감)"이
  같은 방향이면 "추세 일치", 반대면 "추세 역행/상쇄"로 표현해라.
- 숫자는 AVG/OPS 3자리, WAR 2자리, wRC+ 1자리, PA는 정수로 말해라.
```

세부 데이터를 토대로 해당 증감의 이유를 분석, 예측하여 설명할 수 있도록 프롬프트를 작성하였다.

프로토타이핑

선수 정보 입력 화면

예측 및 해설 화면

결론 및 향후 방향

결론

본 프로젝트는 KBO 타자 데이터를 기반으로 다음 시즌 성적을 증감(Δ) 관점에서 예측하는 모델을 구현하고, 이를 시각화 및 해설 기능과 함께 서비스 형태로 구성하는 것을 목표로 진행하였다.
사실 야구 성적 예측은 파크 팩터, 리그 환경 변화, 선수의 컨디션 및 부상, 갑작스러운 반등과 같은 현실적으로 고려하기 어려운 변수가 매우 많은 영역이기 때문에, 모든 상황을 완벽하게 반영하는 모델을 구축하는 데에는 한계가 있다.

그럼에도 불구하고 본 프로젝트에서는 단순히 직전 시즌 성적을 그대로 사용하는 방식이 아니라, 최근 성적 흐름, 연령 효과(에이징 커브), 출장 안정성(PA) 등을 함께 고려함으로써,과도한 낙관이나 비관을 피한 합리적인 방향성의 예측 결과를 확인할 수 있었다.

무엇보다도 본 프로젝트는 “정답을 맞히는 것” 자체보다,
야구라는 도메인을 데이터와 모델링으로 해석해보는 과정의 재미와 의미를 중요하게 두고 진행하였으며,그 과정에서 기대 이상의 결과와 함께 충분한 학습 경험과 흥미를 얻을 수 있었다는 점에서 개인적으로 만족스러운 프로젝트였다.

향후 방향

타자뿐만 아니라 투수 성적 예측 모델을 추가하여, 선발·불펜 등 역할별 특성을 반영한 보다 종합적인 예측 시스템으로 확장할 수 있을 것이다.
또한, 예측 결과를 단순히 LLM을 통해서만 보여주는 것이 아닌 챗봇 기능을 활용하여 예측에 대한 해석을 요청하거나 특정 지표가 왜 그렇게 예측되었는지 추가 질문을 주고받을 수 있는 형태로 확장하고자 한다