[기계학습의이해] 최종 프로젝트 보고서

프로젝트명: MLB 투수 투구 예측 봇 (Pitch Predictor Bot)

학번:202401765

이름: 이혜서

제출일:2025. 12. 10

1. 프로젝트 개요 (Project Overview)

1.1 프로젝트 배경 및 필요성

야구는 투수 놀음이라고 불릴 정도로 투수와 타자 간의 수 싸움이 치열한 스포츠이다. 특히 현대 야구는 데이터 분석이 필수적인 요소로 자리 잡았다. 팬의 입장에서 경기를 관람할 때 현재 볼 카운트와 주자 상황에 따라 투수가 다음에 어떤 공을 던질지 예측하는 것은 경기를 즐기는 또 하나의 큰 재미 요소이다.

이 프로젝트에서는 머신러닝 기법을 활용하여 특정 투수의 과거 데이터를 학습하고, 경기 상황에 따른 다음 투구 구종을 예측하는 AI 모델 및 웹 서비스를 개발하고자 한다.

1.2 목표

- MLB 투수(Logan Allen)의 실제 투구 데이터를 수집 및 분석한다.
- Random Forest 분류 모델을 활용하여 상황별 투구 패턴을 학습한다.
- 사용자가 웹(Streamlit)에서 상황을 입력하면 다음 구종을 예측해주는 서비스를 구현한다.

2. 데이터 수집 및 분석

2.1 데이터 수집

- **데이터 출처:**MLB Statcast 데이터 (Python pybaseball라이브러리 활용)
- **대상 선수:**Logan Allen (Cleveland Guardians, 좌완 투수)
- **수집 기간:**2024년 5월 1일 ~ 5월 3일 경기 데이터
- **데이터 크기:**전처리 전 약 10,000개 이상의 MLB 전체 데이터 중 해당 선수의 데이터 추출

2.2 데이터 전처리

- **Feature (입력 변수):**balls(볼), strikes(스트라이크), on_1b/2b/3b(주자 유무),

아웃 카운트, 이닝

- **Target (예측 변수):**pitch_type(구종)
- **전처리 과정:**

- 결측치(NaN) 처리: 주자 정보가 없는 경우 '주자 없음(0)'으로 보간.
- Label Encoding: 문자열로 된 구종(FF, CH 등)을 수치형 데이터(0, 1...)로 변환.

2.3 데이터 분석 결과 (EDA)

- 해당 선수는 4-Seam Fastball(직구, FF)을 가장 주무기로 사용하며 결정구로 Sweeper(스위퍼, ST)와 Changeup(체인지업, CH)을 섞어 던지는 패턴을 보임.
- 직구와 변화구 간의 구속 차이가 뚜렷하여 타자의 타이밍을 뺏는 전략을 구사함.

3. 머신러닝 모델링

3.1 모델 선정

- **알고리즘:**Random Forest Classifier
- **선정 이유:**야구 데이터와 같은 정형 데이터 분류 문제에서 우수한 성능을 보이며, 과적합 방지에 유리함.

3.2 학습 과정

- **데이터 분할:**전체 데이터를 학습(Train) : 검증(Validation) : 평가(Test) 비율을 약 7 : 1 : 2 로 분할하여 학습의 투명성과 객관성을 확보함.
- **학습 결과:**
- **정확도(Accuracy):**약 47.83%
- **F1-Score:**0.2747

성능 분석:

- 직구(FF)에 대한 예측 정확도는 높으나, 데이터 샘플이 적은 구종(Splitter, Sinker)에 대해서는 예측력이 다소 떨어지는 클래스 불균형 문제가 확인됨.
- 혼동 행렬 분석 결과 모델이 애매한 상황에서는 가장 빈도가 높은 '직구'로 예측하려는 경향(Bias)이 있음을 발견함.

4. 서비스 구현

4.1 시스템 구조

- **Framework:**Python Streamlit라이브러리를 사용하여 별도의 프론트엔드 지식 없이 웹 애플리케이션 구축.
- **구동 방식:**

1. 학습된 모델 파일(pitcher_model.pkl)과 인코더(pitch_type_encoder.pkl)를 로드.
2. 사용자로부터 UI를 통해 볼카운트, 주자 상황 등을 입력받음.
3. 모델이 실시간으로 추론하여 가장 확률이 높은 구종과 확률 분포 그래프를 시각화하여 출력.

5. 실제 사용 결과 (Real Usage)

개발된 모델을 사용하여 실제 경기 상황 5가지를 가정하고 테스트를 진행하였다.

Deploy

⚾ Logan Allen 투구 예측 봇

현재 경기 상황을 입력하면 다음 구종을 예측합니다!

📊 볼카운트 & 이닝

이닝 (Inning)

5 − +

볼 (Balls)

3

스트라이크 (Strikes)

1

아웃 (Outs)

1

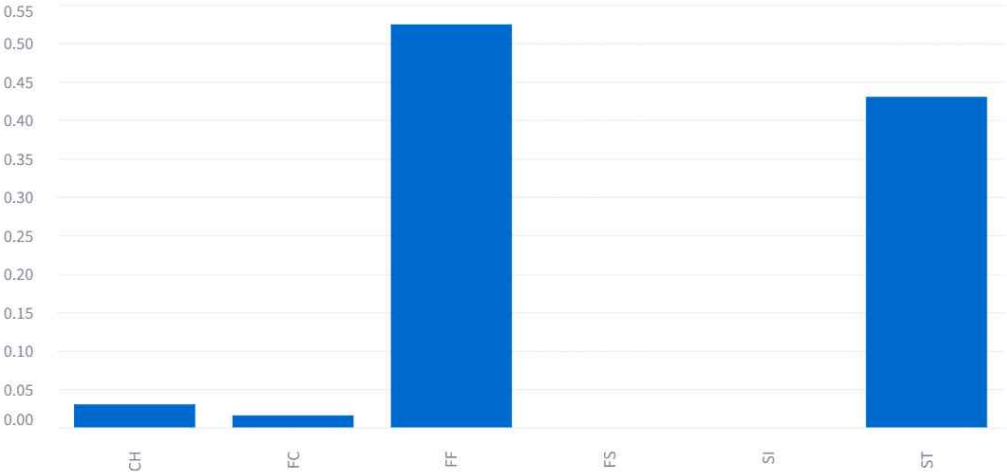
🥎 다음 공 예측하기

🏃 주자 상황

- [x] 1루 주자 있음
- [x] 2루 주자 있음
- [ ] 3루 주자 있음

예측 결과: 직구 (4-Seam Fastball)

📈 구종별 확률 분포



› 상세 확률 보기

Deploy

⚾ Logan Allen 투구 예측 봇

현재 경기 상황을 입력하면 다음 구종을 예측합니다!

📊 볼카운트 & 이닝

이닝 (Inning)

6 − +

볼 (Balls)

3

스트라이크 (Strikes)

1

아웃 (Outs)

1

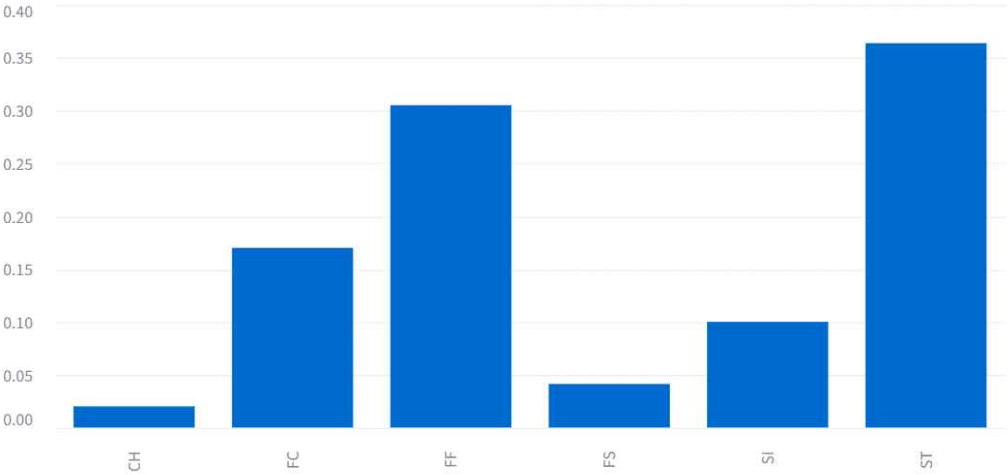
🎯 다음 공 예측하기

🏃 주자 상황

- [x] 1루 주자 있음
- [x] 2루 주자 있음
- [ ] 3루 주자 있음

예측 결과: 스위퍼 (Sweeper) ↔️

📈 구종별 확률 분포



상세 확률 보기

구종	확률
CH	2.00%
FC	17.00%
FF	30.50%
FS	4.12%
SI	10.00%
ST	36.37%

⚾ Logan Allen 투구 예측 봇

현재 경기 상황을 입력하면 다음 구종을 예측합니다!

📊 볼카운트 & 이닝

이닝 (Inning)

8 − +

볼 (Balls)

3

스트라이크 (Strikes)

1

아웃 (Outs)

1

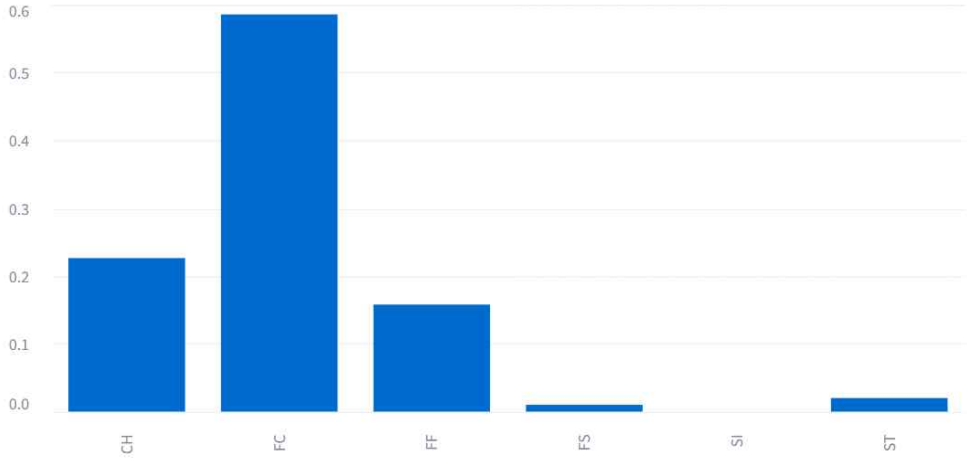
🎯 다음 공 예측하기

🏃 주자 상황

- [x] 1루 주자 있음
- [x] 2루 주자 있음
- [ ] 3루 주자 있음

예측 결과: 커터 (Cutter) 🎀

📈 구종별 확률 분포



> 상세 확률 보기

⚾ Logan Allen 투구 예측 봇

현재 경기 상황을 입력하면 다음 구종을 예측합니다!

📊 볼카운트 & 이닝

이닝 (Inning)

7 − +

볼 (Balls)

0

스트라이크 (Strikes)

1

아웃 (Outs)

1

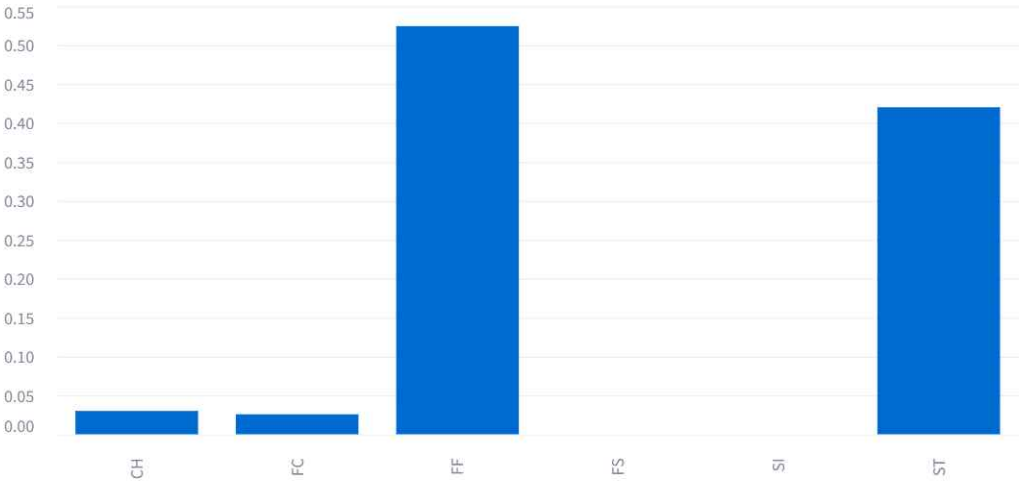
🔮 다음 공 예측하기

🏃 주자 상황

- [x] 1루 주자 있음
- [x] 2루 주자 있음
- [ ] 3루 주자 있음

예측 결과: 직구 (4-Seam Fastball)

📈 구종별 확률 분포



› 상세 확률 보기

Deploy

⚾ Logan Allen 투구 예측 봇

현재 경기 상황을 입력하면 다음 구종을 예측합니다!

📊 볼카운트 & 이닝

이닝 (Inning)

7 − +

볼 (Balls)

3

스트라이크 (Strikes)

1

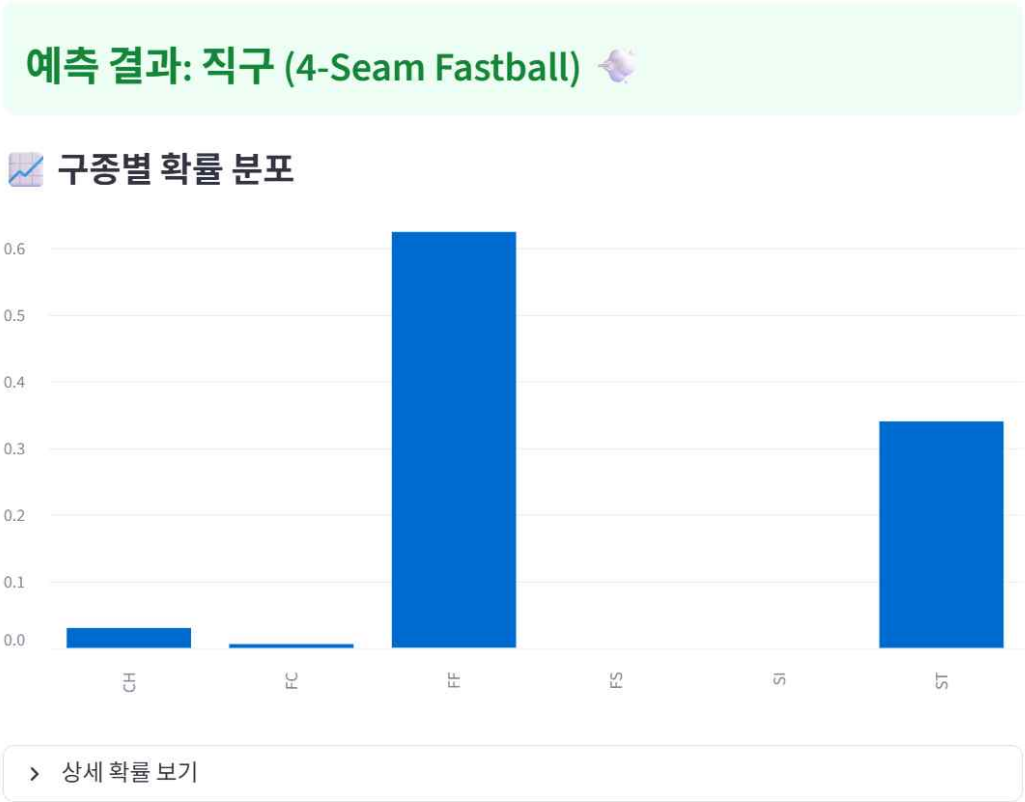
아웃 (Outs)

1

🔮 다음 공 예측하기

🏃 주자 상황

- [x] 1루 주자 있음
- [x] 2루 주자 있음
- [ ] 3루 주자 있음



6. 결론 및 고찰 (Conclusion)

6.1 성과

- 실제 MLB 데이터를 수집부터 전처리, 모델 학습, 그리고 웹 서비스 배포까지 머신러닝 프로젝트의 전 과정을 직접 수행함.
- 단순한 코드 실행을 넘어, GUI 기반의 웹 애플리케이션을 통해 비전문가도 쉽게 예측 결과를 확인할 수 있도록 함.

6.2 한계점 및 개선 방향

- **데이터 부족:**약 100~200개의 데이터로는 투수의 복잡한 심리나 패턴을 완벽히 학습하기에 부족했음. 향후 시즌 전체 데이터를 크롤링하여 학습한다면 정확도를 60% 이상으로 끌어올릴 수 있을 것으로 기대됨.
- **단순한 Feature:**현재는 볼카운트와 주자 상황만 보지만, '상대 타자 정보(좌타/우타)', '점수 차이', '이전 투구 구종' 등의 파생 변수를 추가한다면 성능이 크게 향상될 것같음.

6.3 소감

이번 프로젝트를 통해 "데이터는 거짓말을 하지 않는다"는 것을 배웠으나 동시에 "데이터가 부족하면 편향된 판단을 할 수 있다"는 머신러닝의 한계도 체험할 수 있었다. 앞으로 더 많은 데이터와 고도화된 모델(XGBoost 등)을 적용해보고싶다.