앙상블 학습을 활용한 고교야구선수 성적 예측 모델 개발

작품 소개

프로젝트 개요	본 프로젝트는 앙상블 기법을 활용해 고교 야구 선수의 성적 예측 모델을 개발하는 것을 목표로 함. 아마추어 야구는 경기 수가 제한적이어서 표본 부족 문제와 함께 선수 평가의 어려움이 존재하기에, 체계적인 예측 모델을 통해 스카우터의 선수 평가를 보조하는 도구를 제공하고자 함. 본 연구에서는 Random Forest, LightGBM, XGBoost, Linear Regression 모델을 활용해 투수와 야수의 성적 예측을 수행하고, 교차 검증과 성능 평가지표인 MSE, R^2 Score를 통해 모델을 최적화함. 또한, 잔차 분석을 통해 모델 예측 결과를 검증함.
데이터 처리	• 본 데이터는 KBSA(대한야구소프트볼협회)에서 제공한 2020~2024년 고교 야구 경기 기록을 바탕으로 직접 수집함. 투수 데이터는 평균자책점(ERA)을 목표로 이닝, 피안타, 탈삼진 등 14개 피처로 구성, 야수 데이터는 OPS(출루율+장타율)을 목표로 타율, 타점, 홈런 등 20개 피처로 구성. 총 투수 1,359건, 야수 1,290건의 데이터가 수집되었으며, 표준편차 기반 노이즈 추가 기법을 활용한 데이터 증강을 통해 약 6,500건의 추가 데이터를 확보함.
모델 설계	• 전체적인 모델 설계 과정: < 데이터 수집 - 표준편차 계산 - 노이즈 합산 - 데이터 증강 - 모델 학습 및 평가 > • 모델 설계는 선형회귀와 앙상블 기법(Random Forest, XGBoost, LightGBM)을 비교하여 진행했으며, 특히 XGBoost와 LightGBM이 비선형 패턴 감지와 성능 면에서 우수하다는 점에서 주목함. 더불어 모델의 신뢰성을 검증하기 위해 잔차 분석을 수행하여 예측 오차의 분포를 확인하고 통계적 일관성을 검토함. 이를 통해 데이터 특성에 대한 우연한 성능 향상 가능성을 배제하고 모델의 예측 결과를 신뢰할 수 있음을 확인함.
결과 분석	• **잔차 분석**: 앙상블 기법은 잔차가 정규분포를 따르며 데이터를 잘 학습했음을 보여줌. LightGBM이 XGBoost보다 약간 더 우수한 성능을 보였고, 선형 회귀는 분산이 커 앙상블 기법이 더 나은 성능을 보임. • **성능 평가 분석**: LightGBM은 빠르고 높은 정확도를 보였으며, 야수 성적 예측에서 뛰어난 성능을 발휘. XGBoost는 긴 학습 시간에도 투수 성적 예측에서 더 우수한 결과를 보였음.
추후 활용 방안	• **스카우트 현장**: 다양한 수준의 데이터를 확보해 모델 편향을 줄이고 예측 일반성 향상. 학교명·경기명 입력 시 OPS 상위 선수 검색 기능 제공. • **선수·구단 관리**: 훈련 강도 조절, 부상 예측에 활용. 경기 내외 데이터와 지도자 도메인 지식 결합으로 예측 정밀도 강화. 구단 데이터와 통합해 전력 보강 및 전략 수립 지원. • **스포츠 산업 확장**: KBO 리그 및 글로벌 야구 생태계 혁신에 기여. 야구 외 다양한 스포츠로 모델 확장 가능.

작품 개발자

성명: 최재석 | 배서현 | 최태림 Email: gdhong@skku.edu | dhgil@skku.edu