성균관대학교
소프트웨어융합대학

연구작품

기계 학습을 통한 NBA 선수연봉 예측 모델 구현

김호진
성균관대학교 소프트웨어대학

NBA Player Salary Prediction based on Machine Learning

Kim Ho Jin
Sungkyunkwan University

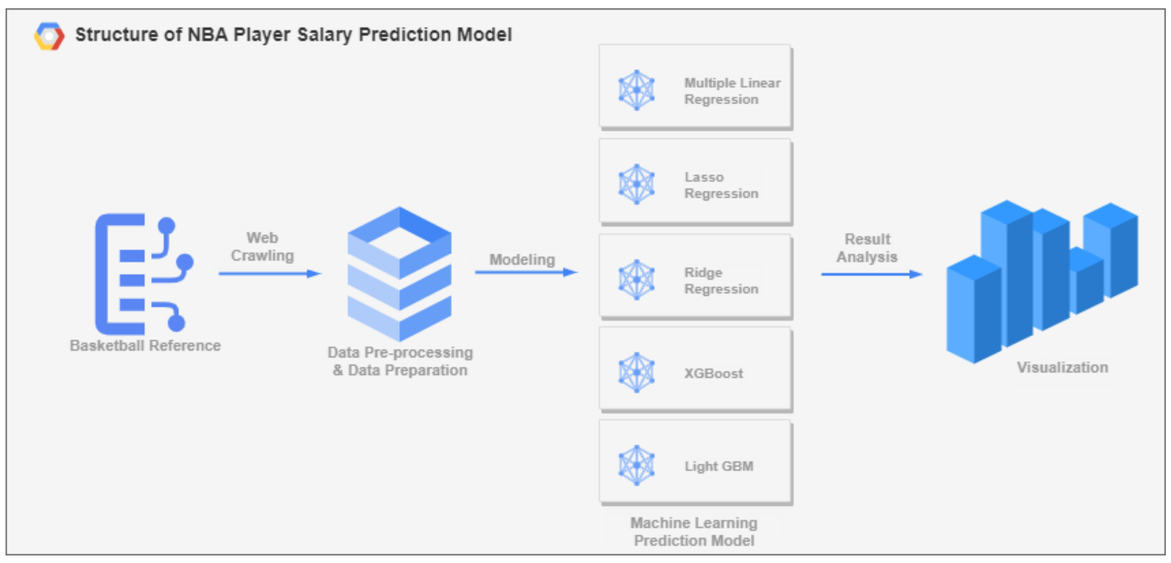
지도 교수: 우홍욱 교수님

연구실명: CSI 연구실

개 요

오늘날 데이터가 산업 전반의 다양한 분야로 활용 폭을 넓히게 되면서 빅데이터를 다루기 위한 기계 학습의 중요성 역시 커지게 되었다. 이러한 추세에 맞춰 스포츠 분야에서도 데이터 활용이 활발하게 이루어지고 있지만, 아직까지 이상적인 데이터셋과 학습 모델이 불분명한 상태이며 개선이 필요한 부분도 상당수 존재한다. 그러므로 본 연구에서는 NBA 선수들의 연봉을 예측하는 모델을 구현하는데 있어서 기존에 사용되었던 1차 기록에 더해 PER, WS, VORP 등의 2차 기록을 함께 사용하고, feature간의 상관관계를 확인한 이후 다중 공선성을 고려하면서 예측 정확도를 높이고자 한다. 해당 모델의 구현은 다양한 종류의 Regression model과 Boosting model을 활용하여 진행한다.

시스템 구성



데이터 수집	NBA의 다양한 통계 정보를 제공하고 있는 Basketball Reference로부터 웹 크롤링을 하여 필요한 데이터들을 얻어온다. 데이터셋은 연봉 관련 데이터와 기록 관련 데이터로 분리하여 가져온 뒤 추후 하나로 병합하여 사용한다.
데이터 준비	수집한 데이터의 feature를 분석하고 이를 기반으로 데이터를 정제한다. 탐색적 데이터 분석은 회귀분석과 히트맵을 이용한 시각화를 통해 진행한다. 회귀 분석은 feature 간의 독립성을 전제로 하기 때문에 예측의 성능을 높이기 위해서 VIF 계수를 고려하여 다중 공선성 문제를 해결한다.
모델 구현	Linear Regression, Lasso Regression, Ridge Regression과 같이 기본적인 기계 학습 알고리즘에 더하여 Boosting 앙상블을 이용하는 XGBoost, LightGBM을 활용해 모델링을 진행한다. 모델 간의 분석 및 비교를 위해 RMSE Score, R-Squared score를 평가지표로 사용한다.
결과 해석	구현 모델 중 가장 정확한 예측도를 보인 모델을 기준으로 Feature importance를 측정하고, 이를 토대로 최종 모델을 구현한다. 최종 모델이 테스트 데이터에 대해 어떻게 동작하는지 확인한 이후 예측 연봉과 실제 연봉 데이터를 비교하는 시각화 자료를 만든다. 해당 자료를 통해서 모델이 얼마나 정확한 예측을 하였는지 직관적으로 확인하고, NBA 선수의 연봉 예측에 있어서 이상적인 데이터셋과 학습 모델을 파악한다.

연구 목표

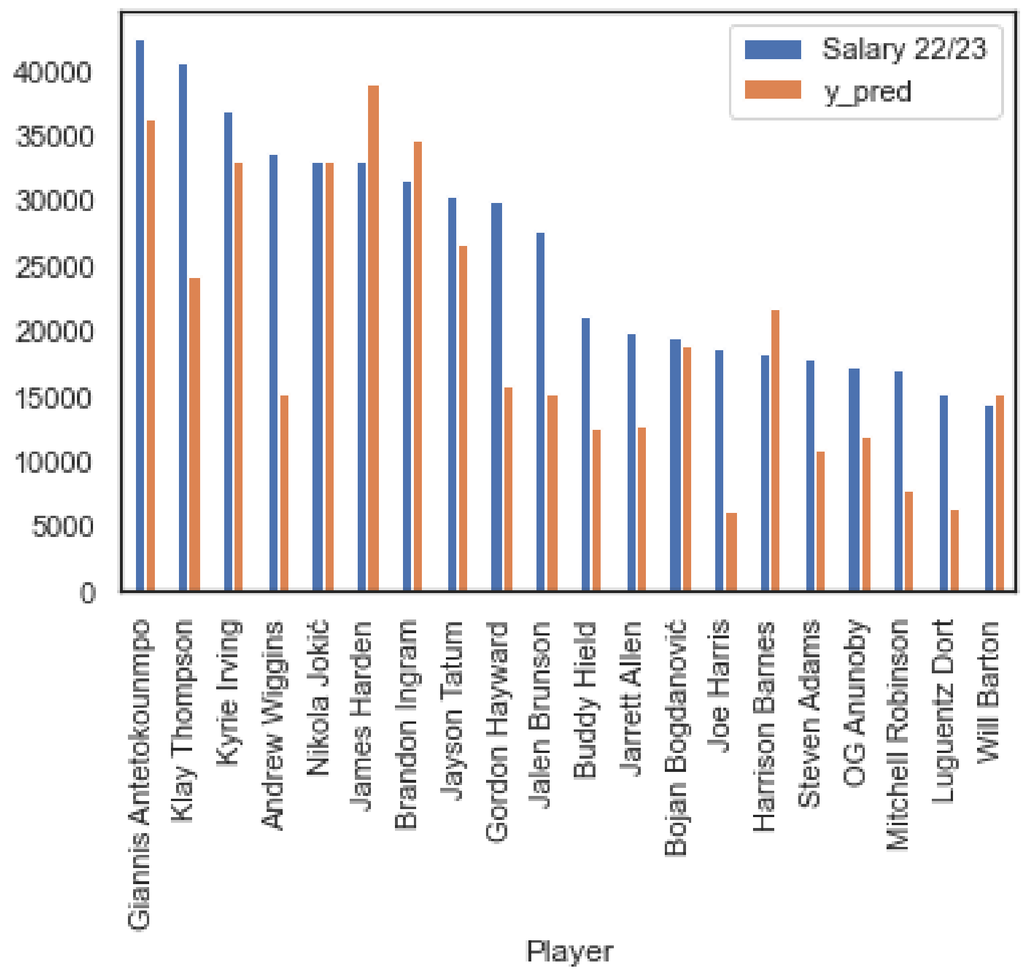
본 연구는 프로스포츠 선수들의 연봉이 선수의 개인 성적과 팀 승리에 대한 기여 등에 영향을 받는다는 가정하에 NBA에서 뛰고 있는 선수의 전년도 성적을 기반으로 다음해 연봉을 예측하는 모델을 만들고자 한다.

결 과

인공지능 모델 별 RMSE / R-Squared 수치는 다음과 같다.

	Linear Regression	Ridge Regression	Lasso Regression	Light GBM	XGBoost
RMSE	4.0602	3.9691	4.349	4.0509	3.8072
R-Squared	0.6549	0.6681	0.6035	0.6556	0.6957

XGB-Regressor가 가장 높은 정확도 점수를 얻었기 때문에 이를 기준으로 최종 모델을 구현했고, 해당 모델이 예측한 연봉과 실제 연봉 데이터를 비교한 시각화 자료는 다음과 같이 나타났다.



결 론

테스트 데이터에 대한 최종 모델의 성능은 Root Mean Squared Error: 3.4604, R-squared Error: 0.7568로 기존보다 향상되었다. 또한, 해당 모델의 예측 값은 실제 연봉과 유사하게 나타났다. 그러나 현재 데이터만으로는 연봉을 산정하는데 있어서 포지션 및 전술상의 차이, 선수들의 부상 여부, 중요 순간 활약도, 기록되지 못하는 공헌도 등을 반영하기 어렵다. 특히 FA 계약과 같은 외부적 요인에 의한 이상치 값은 정확한 연봉 예측에 있어서 큰 혼란을 유발했다. 그러므로 추후 진행되는 연구에서 해당 요소들을 적절하게 조정하여 예측에 반영한다면, 본 연구보다 개선된 결과를 얻을 수 있을 거라 기대한다.