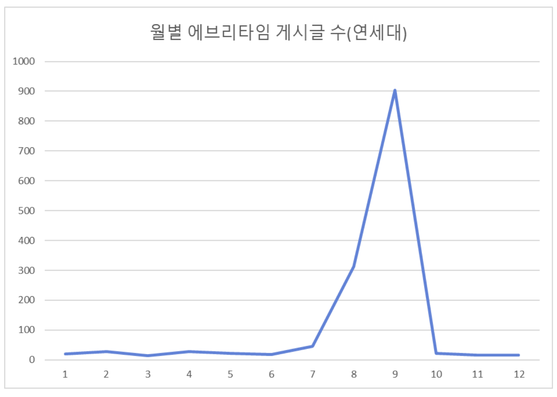
2024 고대 x 연대 산업공학 학술 교류제

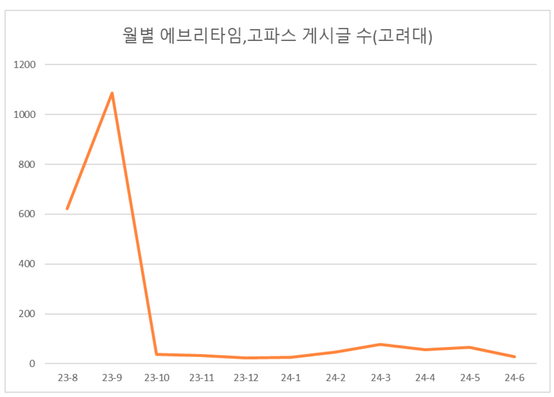
대학 농구 승패 요인 분석 및 2024 정기연고전 농구 경기 결과 예측

데이터분석 1조 : 권해찬, 강수민, 유선호, 전유하, 조영인

Ⅰ. 연구 배경

MZ 대학생들은 대학 생활을 더욱 풍성하게 즐길 수 있는 학교행사, 특히 연고대의 경우 연고전/고연전에 대한 관심을 높은 것으로 나타났다. 연고대의 에타,고파스 및 유튜브 조사 분석 결과 대다수의 학생이 연고전에 큰 기대감을 갖고 있으며, 관련 컨텐츠가 일반 컨텐츠에 비해 매우 높은 조회수를 기록하고 있다. 본 연구는 연고전 관련 데이터 분석을 통해 경기의 승부를 예측하는 정보를 제공함으로써 학생들의 참여와 소통을 촉진하고자 한다. 이를 통해 학생들은 경기 전부터 다양한 예측과 논의로 더욱 재미있고 의미있는 방식으로 연고전을 즐길 수 있을 것이며, 학교에 대한 자부심과 소속감을 강화하는데 기여할 수 있다.





Ⅱ. Data & Methodology

1. 활용 데이터셋

본 연구에서는 2022, 2023, 2024년 KUSF 대학농구 U-리그의 고연전 농구 경기 데이터를 활용하였다. 데이터셋은 각 팀의 경기 결과와 관련된 33개의 피처로 구성되어 있으며, 여기에는 각 쿼터별 스코어, 공격리바운드, 수비리바운드, 2점슛, 3점슛, 자유투 등이 포함된다. 한국대학농구연맹 사이트의 경기 기록지를 통해 데이터를 추출하였으며, 총 299개의 경기를 분석 대상으로 하였다. 각 경기에서 양 팀의 데이터를 각각 활용하여 총 598개의 데이터를 연구에 사용하였다.

2. 실험구조

데이터 전처리와 기본 분석을 통해 원시 데이터를 정리하고 불필요한 값을 제거하여 분석에 적합한 형태로 가공하였다. 이후 여러 분류 모델을 적용해 승리 여부와 관련된 피처들을 추출하고, 이를 통해 경기 결과에 중요한 요소들을 식별하였다. 선별된 피처들을 바탕으로 고연전 경기 결과를 예측할 수 있는 모델을 구축하였으며, 다양한 머신러닝 기법을 활용해 성능을 최적화하여 경기 결과를 정확하게 예측하였다.

3. 방법론

(1) 요인 분석

본 연구에서는 경기 결과에 영향을 미치는 피처를 추출하기 위해 8개의 분류 모델을 활용하고, 고려대와 연세대 경기에 가중치를 부여하여 분석을 진행하였다. 데이터셋은 훈련 데이터와 테스트 데이터를 8:2로 나누고, 하이퍼파라미터 선정에는 그리드 서치 기법을 활용하였다. **상관관계 분석, PCA, SVD, 로지스틱 회귀, 랜덤 포레스트, Gradient Boosting, 라쏘 회귀, 인공신경망** 등의 방법론을 사용하여 모델 성능을 최적화하였다.

(2) 예측모델 구축

고연전 경기 결과 예측을 위해 **로지스틱 회귀, 랜덤 포레스트, Gradient Boosting, 다층 퍼셉트론(MLP), 서포트 벡터 머신(SVM)** 등 다양한 머신러닝 기법을 활용해 총 40개의 예측 모델을 구축하였다. 각 모델에 피처 중요도가 높은 8개의 피처 세트를 적용하고, 그리드 서치를 통해 하이퍼파라미터를 최적화하여 예측 정확성을 극대화하였다.

Ⅲ. Results

본 연구에서는 각 분류 모델별로 평균 이상의 중요도를 가진 피처들을 식별하여 예측 모델의 성능을 높일 수 있는 주요 피처들을 도출하였다.

상관관계 분석에서는 총 득점, 2점슛 성공 개수, 총 리바운드, 어시스트, 1쿼터 득점, 2점슛 성공률, 2쿼터 득점, 3쿼터 득점, 2점슛 시도 개수가 중요한 피처로 선정되었다.

PCA와 SVD 분석에서는 2점슛 성공 개수, 총 득점, 어시스트, 2점슛 시도 개수, 3점슛 성공 개수, 3점슛 성공률, 2점슛 성공률, 3쿼터 득점, 총 리바운드가 중요한 피처로 나타났다.

로지스틱 회귀 모델에서는 총 리바운드, 2점슛 성공률, 스틸, 3점슛 성공 개수, 자유투 성공 개수, 총 득점, 득실차, 어시스트, 2쿼터 득점이 중요한 피처로 선정되었다.

랜덤 포레스트 모델에서는 총 득점, 총 리바운드, 2점슛 성공 개수, 어시스트, 1쿼터 득점이 주요 피처로 나타났다.

Gradient Boosting 모델에서는 총 득점, 총 리바운드, 2점슛 성공 개수, 턴오버, 3점슛 성공률이 중요한 피처로 선정되었다.

라쏘 회귀 모델에서는 총 리바운드, 3점슛 성공 개수, 총 득점, 스틸, 2점슛 성공률, 자유투 성공 개수, 득실차, 어시스트, 블록슛이 중요한 피처로 나타났다.

MLP 분류 모델에서는 총 리바운드, 스틸, 턴오버, 3점슛 시도 개수, 2점슛 성공률이 주요 피처로 선정되었다.

2024 연고전 경기 결과 예측

승리팀		Feature Set						
		Corr	PCA, SVD	LR	RF	GB	Lasso	MLP
Model	LR	Korea	Korea	Korea	Korea	Yonsei	Korea	Korea
	RF	Korea	Korea	Korea	Korea	Yonsei	Korea	Korea
	GB	Korea	Korea	Korea	Korea	Yonsei	Korea	Yonsei
	MLP	Yonsei	Yonsei	Korea	Yonsei	Yonsei	Korea	Korea
	SVM	Korea	Korea	Korea	Korea	Yonsei	Korea	Korea

Ⅳ. Conclusion

본 연구는 고연전 경기 결과 예측을 위해 다양한 머신러닝 모델과 피처 세트를 활용하여 최적의 예측 모델을 구축하였다. 이를 통해 팀의 전략적 의사결정 지원, 선수 성과 평가 및 개선, 대학 농구 리그의 데이터 분석 수준 향상을 기대할 수 있다.

연구의 한계점으로는 데이터의 제한성, 피처 선택의 한계, 모델의 복잡성과 해석 가능성이 있다. 이는 모델의 일반화 성능과 실제 적용 시 이해의 어려움을 초래할 수 있다.

향후 연구에서는 더 많은 데이터를 포함한 분석, 다양한 피처를 고려한 모델링, 해석 가능한 모델 개발, 실시간 데이터 분석 및 예측 시스템 구축이 필요하다. 이를 통해 모델의 성능과 실제 활용 가능성을 높일 수 있을 것이다.